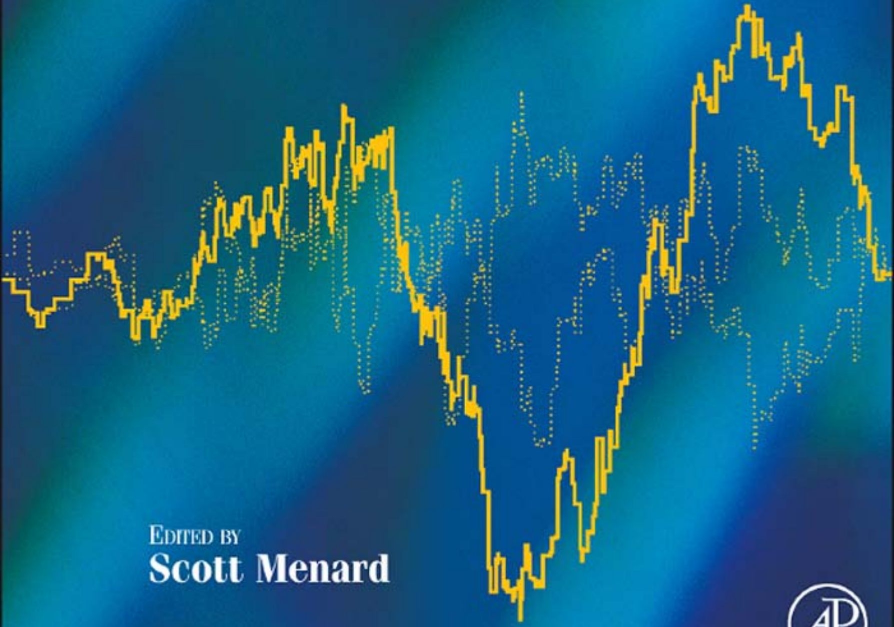
Handbook of Longitudinal Research

Design, Measurement, and Analysis



Edited by

Scott Menard

Handbook of Longitudinal Research: Design, Measurement, and Analysis

*In memory of my parents, Joyce and Marcel Menard, for the values they taught me,
for their help and encouragement in my education, and for always
being there when I needed them most.*

Handbook of Longitudinal Research: Design, Measurement, and Analysis

Editor

Scott Menard,
College of Criminal Justice,
Sam Houston State University,
Huntsville, Texas, USA.

Institute of Behavioral Science,
University of Colorado,
Boulder, Colorado, USA.

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, California 92101-4495, USA
84 Theobald's Road, London WC1X 8RR, UK

First edition 2008

Chapter 10 reprinted from *Journal of Consulting and Clinical Psychology* 61: 911–919, 1993.
Chapter 16 reprinted from *Social Science Research* 2: 133–144, 1977, with permission from Elsevier.
Chapter 36 reprinted from *Analyzing Social and Political Change,* Angela Dale and Richard Davies, eds, SAGE Publications, 1994.

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

Library of Congress Cataloging-in-Publication Data
A catalogue record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 978-0-12-370481-8

For information on all Academic Press publications
visit our web site at www.books.elsevier.com

Printed and bound in the United States of America
08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Contents

List of Contributors ix

Preface xi

Part I Longitudinal Research Design 1

1 Introduction: Longitudinal research design and analysis 3
Scott Menard

2 Using national census data to study change 13
Margo Anderson

3 Repeated cross-sectional research: the general social surveys 33
Tom W. Smith

4 Structuring the National Crime Victim Survey for use in longitudinal analysis 49
Lawrence Hotchkiss and Ronet Bachman

5 The Millennium Cohort Study and mature national birth cohorts in Britain 67
Heather E. Joshi

6 Retrospective longitudinal research: the German Life History Study 85
Karl Ulrich Mayer

Part II Measurement Issues in Longitudinal Research 107

7 Respondent recall 109
Jennifer K. Grotpeter

8 A review and summary of studies on panel conditioning 123
David Cantor

9 Reliability issues in longitudinal research 139
Toon W. Taris

10 Orderly change in a stable world: The antisocial trait as a chimera 153
Gerald R. Patterson

11 Minimizing panel attrition 167
Heather Laurie

12 Nonignorable nonresponse in longitudinal studies 185
E. Michael Foster and Anna Krivelyova

Part III Descriptive and Causal Analysis in Longitudinal Research 197

13 Graphical techniques for exploratory and confirmatory analyses of longitudinal data 199
Garrett M. Fitzmaurice

14 Separating age, period, and cohort effects in developmental and historical research 219
Scott Menard

15 An introduction to pooling cross-sectional and time series data 233
John L. Worrall

16 Dynamic models and cross-sectional data: the consequences of dynamic misspecification 249
Ronald Schoenberg

17 Causal analysis with nonexperimental panel data 259
David F. Greenberg

18 Causal inference in longitudinal experimental research 279
Jos W. R. Twisk

Part IV Description and Measurement of Qualitative Change 295

19 Analyzing longitudinal qualitative observational data 297
Johnny Saldaña

20 Configural frequency analysis of longitudinal data 313
Alexander von Eye and Eun Young Mun

21 Analysis of longitudinal categorical data using optimal scaling techniques 333
Catrien C. J. H. Bijleveld

22 An introduction to latent class analysis 357
C. Mitchell Dayton

23 Latent class models in longitudinal research 373
Jeroen K. Vermunt, Bac Tran and Jay Magidson

Part V Timing of Qualitative Change: Event History Analysis **387**

24 Nonparametric methods for event history data: descriptive measures 389
C. M. Suchindran

25 The Cox proportional hazards model, diagnostics, and extensions 405
Janet M. Box-Steffensmeier and Lyndsey Stanfill

26 Parametric event history analysis: an application to the analysis of recidivism 421
Hee-Jong Joo

27 Discrete-time survival analysis: predicting *whether*, and if so *when*, an event occurs 441
Margaret K. Keiley, Nina C. Martin, Janet Canino, Judith D. Singer and John B. Willett

Part VI Panel Analysis, Structural Equation Models, and Multilevel Models **465**

28 Generalized estimating equations for longitudinal panel analysis 467
Joseph M. Hilbe and James W. Hardin

29 Linear panel analysis 475
Steven E. Finkel

30 Panel analysis with logistic regression 505
Scott Menard

31 Latent growth curve models 523
Michael Stoolmiller

32 Multilevel growth curve analysis for quantitative outcomes 545
Douglas A. Luke

33 Multilevel analysis with categorical outcomes 565
Scott Menard

Part VII Time Series Analysis and Deterministic Dynamic Models **577**

34 A brief introduction to time series analysis 579
Scott Menard

35 Spectral analysis 601
William W.S. Wei

36 Time-series techniques for repeated cross-section data 621
David Sanders and Hugh Ward

37 Differential equation models for longitudinal data 639
Steven M. Boker

38 Nonlinear dynamics, chaos, and catastrophe theory 653
Courtney Brown

Index 665

Contributors

Anderson, Margo History and Urban Studies, University of Wisconsin, Milwaukee

Bachman, Ronet Department of Sociology, University of Delaware

Bijleveld, Catrien C. J. H. NSCR Institute for the Study of Crime and Law Enforcement, Leiden, The Netherlands

Boker, Steven M. Department of Psychology, University of Virginia

Box-Steffensmeier, Janet M. Department of Political Science, The Ohio State University

Brown, Courtney Department of Political Science, Emory University

Canino, Janet Purdue University

Cantor, David Westat and Joint Program for Survey Methodology, University of Maryland

Dayton, C. Mitchell Department of Measurement, Statistics, and Evaluation, University of Maryland

Finkel, Steven E. Department of Policical Science, University of Pittsburgh and Hertie School of Governance, Berlin, Germany

Fitzmaurice, Garret School of Public Health, Harvard University

Foster, E. Michael School of Public Health, University of North Carolina, Chapel Hill

Greenberg, David F. Department of Sociology, New York University

Grotpeter, Jennifer K. Institute of Behavioral Science, University of Colorado, Boulder

Hardin, James W. Department of Epidemiology and Biostatistics, University of South Carolina

Hilbe, Joseph M. Sociology and Statistics, Arizona State University

Hotchkiss, Lawrence Information Technology – User Services, University of Delaware

Joo, Hee-Jong College of Criminal Justice, Sam Houston State University

Joshi, Heather Centre for Longitudinal Studies, Institute of Education, University of London

Keiley, Margaret K. Human Development and Family Studies, Auburn University

Laurie, Heather Institute for Social and Economic Research, University of Essex, UK. Krivelyova, ORC Macro, Inc.

Luke, Douglas A. School of Public Health, St. Louis University

Magidson, Jay Statistical Innovations, Inc.

Martin, Nina Kennedy Center Institute, Vanderbilt University

Mayer, Karl Ulrich Department of Sociology and Center for Research on Inequalities and the Life Course (CIQLE), Yale University

Menard, Scott College of Criminal Justice, Sam Houston State University, and Institute of Behavioral Science, University of Colorado, Boulder

Mun, Eun Young Center for Alcohol Studies, Rutgers University

Patterson, Gerald R. Oregon Social Learning Center

Saldaña, Johnny School of Theatre and Film, Arizona State University

David Sanders Department of Government, Essex University, UK

Schoenberg, Ronald Aptech Systems, Inc. and the University of Washington

Singer, Judith D. Graduate School of Education, Harvard University

Smith, Tom W. National Opinion Research Center and University of Chicago

Stanfill, Lyndsey Department of Political Science, The Ohio State University

Stoolmiller, Michael Oregon Social Learning Center and College of Education, University of Oregon

Suchindran, C. M. Department of Biostatistics, University of North Carolina, Chapel Hill

Taris, Toon W. Work and Organizational Psychology, Radboud University Nijmegen, Nijmegen, The Netherlands

Tran, Bac U. S. Census Bureau

Twisk, Jos. W. R. Department of Clinical Epidemiology and Biostatistics, VU Medical Centre and Institute of Health Sciences, Vrije Universiteit, Amsterdam, The Netherlands

Vermunt, Jeroen K. Department of Methodology and Statistics, Tilburg University

von Eye, Alexander Department of Psychology, Michigan State University

Ward, Hugh Department of Government, Essex University, UK

Wei, William W. S. Department of Statistics, Temple University

Willett, John B. Graduate School of Education, Harvard University

Worrall, John L. Program in Criminology, University of Texas, Dallas

Preface

This handbook is intended to provide a broad, interdisciplinary overview of longitudinal research designs and longitudinal data analysis. Many of the chapters are written at an introductory level, to introduce readers to topics with which they may not be familiar, but some topics, unavoidably, require a higher level of mathematical sophistication than others. Even for those relatively more mathematically challenging chapters, the hope is that the reader (even the novice reader) will gain enough insight into the application and potential of the analytical methods to better understand them when they are presented, and to determine whether they would be helpful in their own research. It is expected that many readers will be relatively well acquainted with some of the methods covered in this volume, but will have much more limited exposure to others. The hope is that for more advanced readers, this handbook may fill in some gaps, and suggest topics worth further examination, particularly by placing them in a context that easily allows comparison with those methods that are more familiar to the reader.

Although this handbook is not primarily designed as a textbook, it can be used as a core text in a course on longitudinal research design, longitudinal data analysis, or a course combining both topics. Sections I and II lay a foundation in longitudinal research design, and along with selected material from other sections, particularly from Section III, would be most suitable for a course emphasizing longitudinal research design. Sections IV–VII focus on different techniques of analysis in longitudinal research, and are perhaps broader in scope than most treatments of longitudinal data analysis. For a survey of longitudinal data analysis, these sections, plus some or all of the chapters from Section III would be appropriate. For a general introduction involving both design and analysis in longitudinal research, this handbook could be used as a core text, supplemented by other books or articles. It is anticipated, however, that this handbook will be most useful as a reference text for practicing researchers in the field of longitudinal research.

This page intentionally left blank

Part I

Longitudinal Research Design

This page intentionally left blank

Chapter 1

Introduction: Longitudinal research design and analysis

Scott Menard

1 Longitudinal and cross-sectional designs for research

As described in Menard (2002), longitudinal research designs can best be understood by contrasting them with cross-sectional research designs. In a purely cross-sectional design, data are collected on one or more variables for a single time period. In longitudinal research, data are collected on one or more variables for two or more time periods, thus allowing at least measurement of change and possibly explanation of change. There are some designs which do not fall neatly under the definition of pure cross-sectional research or longitudinal research. One example is research in which data are collected for different times for different cases, but only once for each variable, and the time dimension is ignored. This design may be used, for example, when data are not all available at the same time, as in Ahluwalia's (1974; 1976) study of economic development and income inequality. Although the data come from more than one time period, the design for any given case, and also the analysis, is cross-sectional in nature. The danger here lies in assuming that relationships are constant over time; the alternative is that any bivariate relationship may reflect not the relationship one would obtain if all of the data were measured for a single period, but may instead be contaminated by changes in that relationship over time.

Another possibility is a time-ordered cross-sectional design, in which each variable is measured only once, but variables are, by design, measured at different times. An example of this is the study by Tolnay and Christenson (1984), who deliberately selected variables which were measured at different times for use in a causal path analysis of fertility, family planning, and development. Each variable was measured at the same time for all countries, but different variables were measured at different times, in order to match the temporal order of measurement with the causal order in the path model. Although measurement occurred at different times for different variables, each variable is measured only once for each case, and the data cannot be used to perform even the simplest true longitudinal analysis (e.g., measuring change in a variable from one period to another). Once again, the design and the analysis are essentially cross-sectional in nature. Had Tolnay and Christenson chosen to postulate instantaneous effects, the analysis could have been performed just as well with purely cross-sectional data.

For the purposes of the analysis (evaluating direct and indirect effects of family planning effort and development on fertility), this design is reasonable, and may have an advantage over models in which causal order in the path model and temporal order of measurement are not the same (Menard and Elliott 1990a). The use of time-ordered cross-sectional data, as in Tolnay and Christenson (1984), is desirable once temporal order has been established, but as described in Menard (2002), it is insufficient to insure that one does not "predict" a cause from its effect. With a true longitudinal design and analysis, it might be possible to ascertain the true causal direction in the relationship between X and Y. With cross-sectional data, even time-ordered cross-sectional data, we run the risk of undetectable misspecification because of incorrect causal ordering in the model being estimated. With longitudinal data, incorrect causal ordering is more likely be detected, and the model can be corrected.

2 Designs for longitudinal research

Menard (2002) describes four basic designs for longitudinal research: total population designs, repeated cross-sectional designs, revolving panel designs, and longitudinal panel designs. These designs are illustrated in Figure 1.1, and examples of each are provided in Chapters 2–6 of Section I in this volume. In Figure 1.1, the horizontal dimension represents the period (a month, year, or decade) for which data are collected, and the vertical dimension represents the cases (population or sample) for which data are collected. Moving from left to right, vertical lines on the left indicate entry into the population or sample being analyzed, and vertical lines on the right indicate exit, as indicated in the first part of Figure 1.1.

In a *total population design*, the total population is surveyed or measured in each period of the study. Because some individuals die and others are born from one period to the next, the cases are not identical from one period to the next, but if the periods are short, the overwhelming majority of cases may be the same from one period to the next. As one example, the decennial census of the United States attempts to collect data on age, sex, ethnicity, and residence of the total population of the United States every ten years, and does so with an accuracy estimated at 95–99% (Hogan and Robinson 2000; Robey 1989). With somewhat lower, but still substantial accuracy and completeness of coverage, the Federal Bureau of Investigation's *Uniform Crime Reports* attempt to collect data on arrests for specific offenses and, for a limited set of offenses, crimes known to the police, plus the age, sex, race, and residence (urban, suburban, or rural) of arrestees for all police jurisdictions in the United States. In Chapter 2 of this volume, Margo Anderson illustrates the use of total population data, specifically census data, for longitudinal research. To the extent that individual data across time are recoverable from the total population data, the total population design permits the use of all possible methods of longitudinal data analysis, but total population designs are most commonly used in aggregate rather than individual level research, and more often involve analytical techniques such as those in Chapters 13–14 (analyzing developmental and historical change) and Section VII (time series analysis and deterministic dynamic models), rather than techniques better adapted to analysis of change at the individual level. In addition to this type of analysis, which focuses on *changes in the values of variables* (e.g., changes in per capita gross national product or changes in homicide rates) over time, this type of design is also well suited to the analysis of *changes in relationships among variables* (e.g., the correlation between ethnicity and political affiliation, or between education and income) over time.

Each of the other three longitudinal designs in Figure 1.1 involves a sample drawn from the total population, and is thus a subset of the total population design. The three designs differ in the extent to which the same or comparable cases are studied from one period to the next.

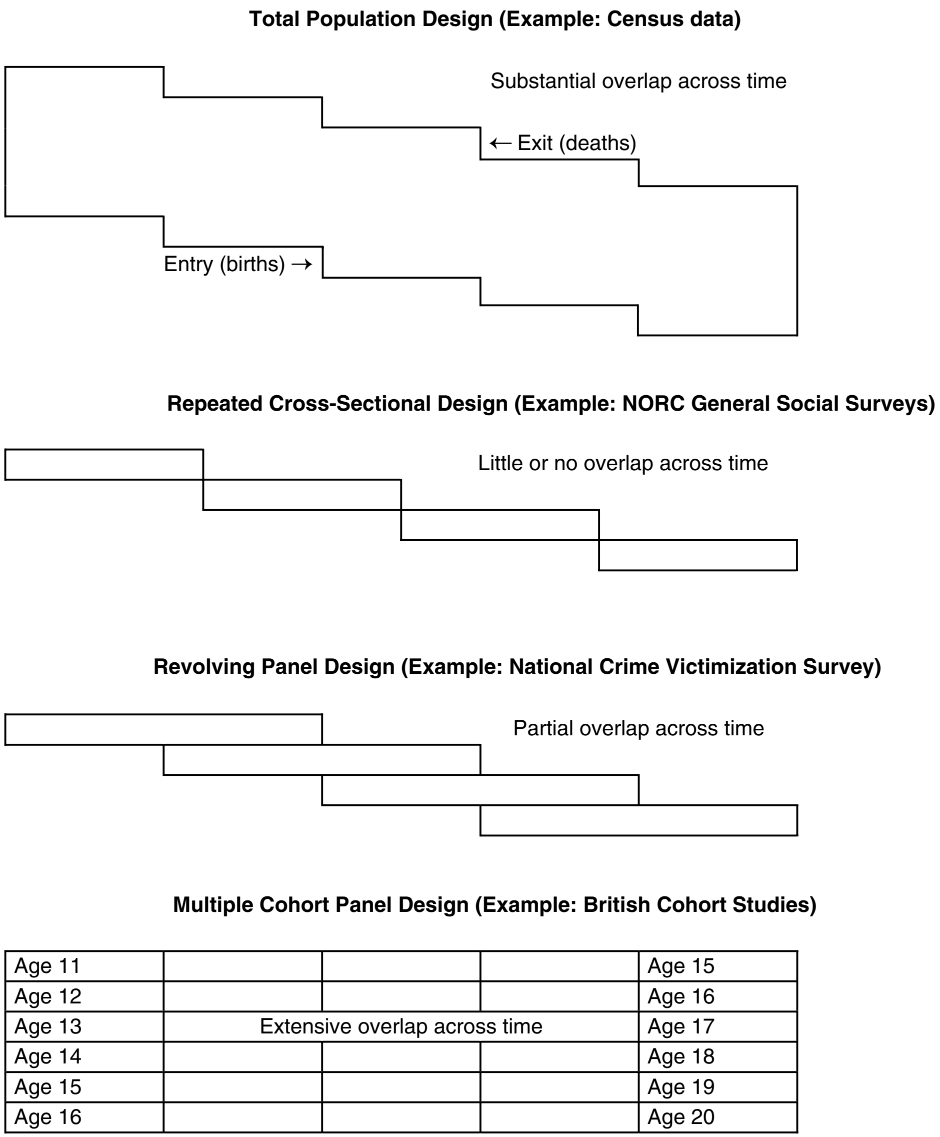

Total Population Design (Example: Census data)
Substantial overlap across time
← Exit (deaths)
Entry (births) →
Repeated Cross-Sectional Design (Example: NORC General Social Surveys)
Little or no overlap across time
Revolving Panel Design (Example: National Crime Victimization Survey)
Partial overlap across time
Multiple Cohort Panel Design (Example: British Cohort Studies)
Age 11
Age 12
Age 13
Age 14
Age 15
Age 16
Extensive overlap across time
Age 15
Age 16
Age 17
Age 18
Age 19
Age 20

Figure 1.1 Longitudinal designs for data collection

This distinction has important implications for which types of longitudinal analysis are possible with each design. In the *repeated cross-sectional design*, the researcher typically draws independent probability samples at each measurement period. These samples will typically contain entirely different sets of cases for each period, or the overlap will be so small as to be considered negligible, but the cases should be as comparable from one period to another as would be the case in a total population design. An example of the repeated cross-sectional design is the General Social Surveys (GSS), which include an annual general population sample survey conducted by the National Opinion Research Center, which covers a wide range of topics, and emphasizes exact replication of questions to permit comparisons across time (Davis and Smith 1992). Thomas W. Smith, in Chapter 3, describes the GSS, including the methods used to collect the data, and gives an overview of the types of research that have been done using this extensive dataset. Much of the research involving repeated cross-sectional data is cross-sectional in nature, and even more than is the case with total population data, the analysis of change in repeated cross-sectional data may involve aggregate level research; and like the total population design, the repeated cross-sectional design is well suited to examine changes in values of variables and in relationships among variables over time.

Revolving panel designs collect data on a sample of cases either retrospectively or prospectively for some sequence of measurement periods, then drop some subjects and replace them with new subjects. The revolving panel design may reduce problems of panel mortality and repeated measurement in prospective studies (to be discussed in Section II), or problems of extended recall periods in retrospective studies. Retention of a particular set of cases over several measurement periods allows short-term measurement of change on the individual or case level, short-term analysis of intracohort developmental change, and panel analysis. Replacement of the subsample which is dropped in a measurement period with a new but comparable subsample of cases permits analysis of long-term patterns of aggregate change, similar to the analyses possible with total population and repeated cross-sectional designs. If the time lag between cause and effect is smaller than the time (periods) for which cases are retained in the sample, analysis of temporal and causal order is possible. The combination of longitudinal data involving repeated measurement on some cases with data which do not involve repeated measurement on others may permit comparisons which can indicate whether repeated measurement is producing any bias in the data (e.g., increased or decreased willingness to report events after either building up some level of trust or finding out that reporting leads to long and tedious follow-up questions). A good example of a revolving panel design is the National Crime Victimization Survey, whose use in longitudinal research is described by Lawrence Hotchkiss and Ronet Bachman in Chapter 4.

In a longitudinal panel design, the same set of cases is used in each period. In practice, there may be some variation from one period to another as a result of missing data. For example, when cases are individuals, some of those individuals may die between one measurement period and the next, others may not agree to cooperate, and others may move to new locations and not be found by the researcher. All of these are sources of *panel attrition*, and apply primarily to *prospective* panel designs, in which measurement or data collection occurs *during* more than one period as well as *for* more than one period. The combination of measurement during more than one period and for more than one period represents, for some scholars, the only true longitudinal design, the only design that allows the measurement and analysis of intraindividual changes in cognitive and behavioral characteristics of individuals. The prospective panel design is here illustrated by Heather Joshi, using examples drawn from British longitudinal cohort studies, in Chapter 5. For this design, the techniques presented in Sections III–VI of this volume, but not Section VII, are generally appropriate.

The analytical methods in Sections III–VI are also appropriate for the analysis of *retrospective* panel designs, in which data collection may occur only once, at a single period, but the data are collected for two or more periods (prior to or during the period in which the data are being collected). In retrospective panel designs, there may be sampling bias as a result of excluding respondents who have died by the last period *for* which the data are collected (or by the time *at* which the data are collected), or from whom data would have otherwise been available for earlier periods but not for the last period. In both retrospective and prospective panel designs, missing data may result from failure of the respondent to remember past events, behaviors, or attitudes, or from unwillingness by the respondent to divulge some information, and also from inability of the researcher to locate or obtain cooperation from some respondents. In principle, there need be no difference in the quality of the data obtained in prospective and retrospective panel designs, although such differences have often been observed in practice. An example of a retrospective panel design with extensive attention to potential issues of data quality is presented in Chapter 6 by Karl Ulrich Mayer, using the German Life History Study (GLHS).

As noted in Menard (2002), the designs diagrammed in Figure 1.1 are not the only possible designs for longitudinal research. It is possible, for example, to have a revolving sample in which subsamples may be dropped for one period, then re-included in the sample in a subsequent period. It is also possible to have a panel design in which cases are dropped, without replacement, after they meet some criterion (e.g., age 21). This latter design would result in a monotonically decreasing sample size which could pose problems for analysis of data from later years of the study (unless the design were further modified by replenishing the sample with new respondents from younger cohorts). The general considerations associated with the various designs for data collection do not change, however, with modifications of the four designs presented in Figure 1.1, and variations on these basic designs must be evaluated in terms of their adequacy for describing short- and long-term historical trends (period effects), intercohort and intracohort developmental changes (age effects), separating age, period, and cohort effects, and ascertaining not only the strength but also the direction of causal influences. Total population designs can, in principle, be used for practically any type of longitudinal analysis, given a sufficient number of cases and measurement periods. Other designs are more limited, and their appropriateness must be judged in the context of a particular research problem.

3 Measurement issues in longitudinal research

Longitudinal research is subject to all of the concerns about measurement that arise in cross-sectional research, plus some issues with particular relevance to longitudinal research. Put another way, longitudinal research has all of the problems of cross-sectional research, plus a few more. In the second section of this handbook, the focus is on those issues most specifically relevant to longitudinal research. Skipping ahead for a moment, in Chapter 9 Toon W. Taris discusses reliability issues in longitudinal research. Taris examines issues of distinguishing unreliability from true change, and raises (not for the last time in this volume) the issue of the reliability of change scores as measures of change. This is followed in Chapter 10 by Patterson's discussion of one of the challenging issues in long-term longitudinal research on individual change, the possibility that it may be appropriate to operationalize the same concept in different ways across the life course. The issue here is that, on one hand, whenever we change the way we measure a concept in longitudinal research, if there appears to be a change, we cannot be certain whether the change results from change in the concept we are trying to measure, or change in the measurement of the concept. Yet for research on individuals over the life course, the same measurement at different stages of the life course may not be validly measuring the same concept because different measures are appropriate at different ages. In much of longitudinal research, there is an emphasis on consistency of measurement, avoiding changes in how a concept is measured because otherwise we cannot tell whether an apparent change represents a true change in the underlying concept or merely in the measurement itself. As Patterson explains, however, using the same operationalization of the same concept over the life course may not always be the best approach, and the same underlying concept may manifest itself, and thus need to be measured, in different ways at different stages of the life course. Taken together, the chapters by Taris and Peterson address the issue of distinguishing true change and stability from measurement effects that mimic change in longitudinal research.

The chapters by Taris and Patterson apply to longitudinal research in general, whether measurement is done prospectively or retrospectively. The remaining chapters deal with issues more specific to different types of longitudinal research. Jennifer Grotpeter in Chapter 7 provides a general conceptual framework for understanding long-term retrospective recall, and examines the results of studies of recall as it is related to the length of the recall period. On this topic, see also Chapter 6 on the (retrospective) German Life History Study in the previous section, in which Karl Ulrich Mayer describes the techniques (and their results) used to enhance recall in a major retrospective panel study. Chapter 8 by David Cantor examines an issue specific to prospective longitudinal research, the effect of panel conditioning in panel research. Panel conditioning potentially occurs when respondents react to previous experience of participating in the study by changing their behavior or answers, possibly in response to their perceptions of what the researcher is seeking, or possibly to reduce their own burden as respondents. Consideration of issues specific to prospective longitudinal panel research continues in Chapter 11 by Heather Laurie, who discusses procedures for minimizing panel attrition in longitudinal samples. Despite our best attempts to minimize panel attrition, however, circumstances beyond our control (and sometimes beyond the control of our respondents) may result in missing data in longitudinal designs. In Chapter 12, E. Michael Foster and Anna Krivelyova present a brief discussion of different types of missing data, along with an example of how to handle nonignorable nonresponse in longitudinal research designs.

4 Descriptive and causal analysis in longitudinal research

The first stage in the process of analyzing longitudinal data is to provide a basic description of the data. The chapters in Section III present issues and techniques which cut across different types of longitudinal research designs. In Chapter 13, Garrett Fitzmaurice describes graphical techniques for presenting longitudinal data. Fitzmaurice shows how exploratory graphical techniques in longitudinal research help in providing insights prior to estimation of the model, and are also useful in the post-estimation diagnostic phase for examining residuals. In Chapter 14, I review the distinction between historical and developmental change and the issues involved in separating the two, with special attention to the disentangling of age, period, and cohort effects. In Chapter 15, John L. Worrall provides an introduction to pooling cross-sectional and time series data, a topic which will recur in other chapters in this handbook. In Chapter 16, Ronald Schoenberg describes the consequences of dynamic misspecification in the use of cross-sectional data to model dynamic processes. Schoenberg's chapter indicates the conditions under which cross-sectional data may be adequate to model dynamic processes, and indicates the consequences of using cross-sectional data when those conditions are not met. In Chapter 17, David Greenberg reviews attempts to draw causal inferences from nonexperimental panel data, tracing the evolution of causal inference in longitudinal research from some of the earliest methodological attempts to more contemporary approaches. Jos W. R. Twisk in Chapter 18 provides a parallel consideration of techniques for drawing causal inferences in longitudinal experimental research.

5 Description and measurement of qualitative change

The definitions of qualitative data and qualitative change may be approached from different perspectives, including how the data were collected, and at what level of measurement (nominal or at most ordinal for qualitative data). While consideration of qualitative data is not excluded from Section III, the focus is on techniques for the presentation and analysis of quantitative data. Section IV begins with Chapter 19, in which Johnny Saldaña describes an approach to the description and measurement of qualitative change in qualitative observational research. Saldaña offers a systematic approach to organizing and analyzing data from qualitative research with an emphasis on tracing patterns of change in qualitative data. Turning from qualitative defined in terms of method to qualitative defined in terms of level of measurement, Alexander von Eye and Eun Young Mun in Chapter 20 describe the use of configural frequency analysis for describing and analyzing qualitative change in longitudinal data. In configural frequency analysis, the emphasis is on tracing change in nominal variables across multiple measurement periods to identify normative and exceptional patterns of change. In Chapter 21, Catrien C. J. H. Bijleveld describes the use of optimal scaling techniques, typically calculated using alternating least squares (ALS) estimation, as a way of "quantifying" qualitative variables, and the applications of optimal scaling to the study of change in longitudinal research. The approaches described by Saldaña, von Eye and Mun, and Bijleveld are perhaps less well known, and typically less well covered, than other techniques for longitudinal data analysis. More widespread at present, at least in the social and behavioral sciences, is the use of latent class analysis to identify different qualitative "types" of individuals or of patterns of behavioral or attitudinal change over time. In a companion pair of chapters, C. Mitchell Dayton in Chapter 22 provides an introduction to latent class analysis, and Jeroen Vermunt, Bac Tran, and Jay Magidson in Chapter 23 describe the application of latent class models in longitudinal research. Taken together, the chapters in Section III offer an array of options for the analysis of data that are qualitative in terms of the research design, the level of measurement, and the assignment of cases to latent qualitative classes in longitudinal research.

6 Timing of qualitative change: event history analysis

Event history analysis is not so much a single technique as a set of related techniques for describing, analyzing, and predicting the timing of qualitative change (including whether it occurs at all). Section V begins with Chapter 24 by C.M. Suchindran, in which the most basic models for event history analysis, life table models for change, are described. These models make minimal distributional assumptions, and hence can be described as distribution free or nonparametric methods. In Chapter 25, Janet M. Box-Steffensmeier and Lyndsey Stanfill describe the Cox proportional hazards model, a semiparametric technique for event history analysis. Parametric event history analysis is briefly described and illustrated by Hee-Jong Joo in Chapter 26. The proportional hazards and parametric event history analysis models both assume that measurements occur fairly continuously in time. This, however, is not the case in much social science research, which may consist of measurements separated by a year or more. For these longer measurement intervals, discrete time event history analysis, as described by Margaret K. Keiley, Nina C. Martin, Janet Canino, Judith D. Singer, and John B. Willett, allows for the occurrence of many events within a single discrete time period. Like parametric event history analysis, discrete time event history analysis makes certain distributional assumptions regarding the parameters in the model. In contrast to the continuous time parametric and semiparametric approaches, discrete time event history analysis works more easily with time-varying covariates and with multiple events occurring in a single time interval, and it can be implemented using ordinary logistic regression or related (e.g., complementary log-log regression) techniques.

7 Panel analysis, structural equation models, and multilevel models

The statistical techniques in Section VI are techniques primarily oriented to the analysis of longitudinal panel data, and would probably be considered by some to be the most mainstream longitudinal analysis methods. The section begins with a discussion by Joseph M. Hilbe and James W. Hardin in Chapter 28 of the generalized estimating equation (GEE) approach to the analysis of longitudinal data. The use of GEE involves the estimation of parameters and standard errors that avoids unrealistic assumptions of independence of observations in longitudinal analysis and adjusts for the dependencies in the data. In Chapter 29, Steven E. Finkel describes approaches to linear panel analysis with quantitative (interval and ratio scaled) outcome variables, and in the following chapter, Chapter 30, I describe the use of linear panel analysis for the analysis of categorical (dichotomous, polytomous nominal, and polytomous ordinal) dependent variables, including the critical issue of how to measure and model change in categorical variables in linear panel models. Taken together with the chapters by Worrall (15), Greenberg (17), Twisk (18), and Hilbe and Hardin (28), these chapters provide an overview of the analysis of short-term quantitative and qualitative change and causal inferences, in which the specific nature of the trajectories or patterns of change is typically not itself being modeled.

The next three chapters turn to the modeling of trajectories of change, usually over the relatively short term, but potentially involving long-term trajectories as well. Michael Stoolmiller in Chapter 31 describes the latent growth curve modeling technique, based on structural equation modeling techniques. Latent growth curve models view trajectories or patterns of change over time as unobserved variables to be treated as latent variables in structural equation modeling. In contrast, in multilevel growth curve analysis of quantitative outcomes, as described by Douglas A. Luke in Chapter 32, one typically attempts to fit a manifest (not latent) polynomial or other function to the data to describe the trajectory of individual cases over time, and to explain variations in those trajectories using a combination of time-invariant individual case characteristics and time-varying covariates. When the dependent variable in the multilevel analysis is categorical rather than quantitative, it may be appropriate to speak not of "growth" curve analysis, but of multilevel change analysis. My focus in Chapter 33 is on showing the application of the logistic regression framework to the multilevel analysis of change, and on highlighting some of the contrasts of multilevel change analysis for categorical dependent variables from multilevel growth curve models for quantitative dependent variables, from event history analysis, and from linear and logistic regression panel analysis.

8 Time series analysis and deterministic dynamic models

Time series analysis stands out from the other methods of analysis in this handbook in the number of cases and the number of time periods. Most often, time series analysis is applied to aggregated data for a single case (a nation, city, corporation, or other aggregate entity, not an individual), or perhaps a handful of such cases, typically analyzed separately rather than together, as in other methods covered in this handbook, and the number of time periods is typically large, often over 100. In Chapter 34 I provide a brief introduction to time series analysis from the perspective of longitudinal research, which is a little different from the perspective out of which time series analysis itself has grown. Here, time series analysis is viewed as one tool for longitudinal research, with more of a focus on description and explanation and less of a focus on forecasting than is typical in the mainstream time series analysis literature. In Chapter 35, William W. S. Wei provides an introduction to spectral analysis, the most mathematically demanding of the time series analysis approaches. In Chapter 36, David Sanders and Hugh Ward provide further details on alternative approaches to time series analysis with one or more predictors included in the model, and offer a useful comparison of the different approaches to time series analysis to the empirical study of public opinion in political science.

The final two chapters in Section VII also involve a higher level of mathematical sophistication than most of the other chapters in this handbook. Steven M. Boker in Chapter 37 describes the application and estimation of differential equation models in longitudinal research using a latent variable structural equation modeling approach to estimate the parameters of the differential equation model. Finally, Courtney Brown in Chapter 38 provides a brief introduction to the application of nonlinear dynamics, chaos, and catastrophe theory to the study of change.

9 Conclusion

One of the goals of this handbook is to make the reader aware of the richness and breadth of research design and analytical techniques available for longitudinal research. The first section of this handbook begins with strong examples of each of the major types of longitudinal research design. Section II focuses on measurement issues that arise in longitudinal research generally, and also more specifically in particular types of longitudinal research designs. With each of these designs, the number of cases and periods may vary, and as a result of this variation, different methods of analysis may be appropriate. The number of cases is in principle independent of the type of design. In a total population design, for example, at the individual level, the total population of a tribal society may number fewer than 100. In aggregate analysis, a cohort or a population, rather than its individual members, may be the unit of analysis, and the number of these aggregate units may be small. At the other end of the continuum, the revolving sample in the National Crime Victimization Survey includes over 100,000 individuals from 60,000 households. All of these possible combinations of type of design and number of cases are included within the broad category of longitudinal research.

The number of cases and the number of time periods, in turn, drives the choice of analytical methods. With no more than a handful of cases but many time periods, the time series and deterministic dynamic models in Section VII are most appropriate. With no more than a handful of time periods but many cases, panel analytic techniques described in Section VI may be best, and as the number of time periods increases up to ten or so, techniques such as latent and multilevel growth curve and change models in Section VI, event history analysis in Section V, and the techniques for qualitative data analysis in Section IV become increasingly feasible. In the best of all possible worlds for longitudinal research, many cases and many time periods, event history and multilevel growth curve and change models seem at present to offer the best options. It is hoped that, by presenting in some detail the different designs for longitudinal research, issues in longitudinal research design, and techniques of analysis for longitudinal data, all in a single sourcebook, readers will be increasingly aware of and better able to make informed selections among the different options available to best capitalize on the strengths of longitudinal research.

References

Ahluwalia, M. S. (1974). Income inequality: Some dimensions of the problem. In H. B. Chenery, M. S. Ahluwalia, C. L. G. Bell, J. H. Duloy and R. Jolly (eds), *Redistribution with Growth: An Approach to Policy* pp. 3–37. Oxford, UK: Oxford University Press.

Ahluwalia, M. S. (1976). Inequality, poverty, and development. *Journal of Development Economics*, 2: 307–342.

Davis, J. A. and Smith, T. W. (1992). *The NORC General Social Survey: A User's Guide*. Newbury Park, CA: Sage.

Hogan, H. and Robinson, G. (2000). *What the Census Bureau's Coverage Evaluation Programs Tell Us About the Differential Undercount*. Washington, DC: US Bureau of the Census.

Menard, S. (2002). *Longitudinal Research*. Thousand Oaks, CA: Sage.

Menard, S. and Elliott, D. S. (1990). Longitudinal and cross-sectional data collection and analysis in the study of crime and delinquency. *Justice Quarterly*, 2: 11–55.

Robey, B. (1989). Two hundred years and counting: The 1990 census. *Population Bulletin*, 4(1): 1–43.

Tolnay, S. E. and Christenson, R. L. (1984). The effects of social setting and family planning programs on recent fertility declines in developing countries: A reassessment. *Sociology and Social Research*, 62: 72–89.

Chapter 2

Using national census data to study change

Margo Anderson

1 Introduction

A census is generally defined as an official count conducted by a national government of a country's population, economic activity or other national phenomena, such as religious institutions. A population census determines the size of a country's population and the characteristics of its people, such as their age, sex, ethnic background, marital status, and income. An economic census collects information on the number and characteristics of farms, factories, mines, or businesses. As state sponsored data collections, censuses are primarily designed for governmental and public policy use. The academic research community uses census data for grounding much social science research, including providing sampling frames for periodic surveys, for classification schemes and standard and consistent questionnaire design and wording, and for comparative analysis, both spatial and temporal. Censuses are official, public, repeated, infrequent, and comprehensive sources of reliable and relatively simple information on a society and also serve as a fundamental longitudinal data source for social science.

Most countries of the world currently conduct population censuses at regular intervals. By comparing the results of successive censuses, analysts can see whether the population is growing, stable, or declining, both in the country as a whole and in particular geographic regions. They can also identify general trends in the characteristics of the population. Censuses are "complete counts" not samples of the phenomenon under study, and accordingly they are very expensive. They require elaborate administrative operations and thus are conducted relatively infrequently. The United States, for example, conducts a population census every ten years (a decennial census), and Canada conducts one every five years (a quinquennial census). Economic censuses are generally conducted on a different schedule from the population census, generally every five years.

Censuses of population usually try to count everyone in the country as of a fixed date, often known as Census Day. Generally, governments collect the information by sending a questionnaire in the mail or a census taker to every household or residential address in the country. The recipients are instructed to complete the questionnaire and send it back to the government, which processes the answers. Trained interviewers visit households that do not respond to the questionnaire and individuals without mail service, such as the homeless or those living in remote areas.

Censuses require significant public cooperation for their operational success since the goal of the census is a snapshot or cross-sectional information collected at a point in time. The responding public needs to know the census is coming, be aware of the responsibility to respond accurately and promptly, and be willing to cooperate. Census questions are designed to be simple and intelligible for the entire population under study. For the society in question, the questions can not be controversial or ambiguous, or the quality of the responses deteriorates rapidly.

2 Official uses of census information

Governments use census information in almost all aspects of public policy. In some countries, the population census is used to determine the number of representatives each area within the country is legally entitled to elect to the national legislature. The Constitution of the United States, for example, provides that seats in the House of Representatives should be apportioned to the states according to the number of their inhabitants. Each decade, the US Congress uses the population count to determine how many seats each state should have in the House and in the electoral college, the body that nominally elects the president and vice president of the United States. This process is known as reapportionment. States frequently use population census figures as a basis for allocating delegates to the state legislatures and for redrawing district boundaries for seats in the House, in state legislatures, and in local legislative districts. In Canada, census population data are similarly used to apportion seats among the provinces and territories in the House of Commons and to draw electoral districts.

Governments at all levels—such as cities, counties, provinces, and states—find population census information of great value in planning public services because the census tells how many people of each age live in different areas. These governments use census data to determine how many children an educational system must serve, to allocate funds for public buildings such as schools and libraries, and to plan public transportation systems. They can also determine the best locations for new roads, bridges, police departments, fire departments, and services for the elderly, children, or the disabled.

3 Public and research use of census information

The official uses of census information do not exhaust the use of the data. Private researchers, including businesses and marketing organizations and the media analyze population and economic census data to determine where to locate new factories, shopping malls, or banks; to decide where to advertise particular products; or to compare their own production or sales against the rest of their industry. Community organizations use census information to develop social service programs and service centers. Censuses make a huge variety of general statistical information about society available to researchers, journalists, educators, and the general public.

In addition to these immediate uses, the academic research community makes fundamental use of census data for grounding much social science research, including providing sampling frames for periodic surveys, for classification schemes and standard and consistent questionnaire design and wording, and for comparative analysis, both spatial and temporal. The national census, in short, is a fundamental building block for social science research. Its particular character as an official, public, repeated, infrequent, and comprehensive source of reliable and relatively simple information on a society in turn shapes both the research uses to which it can be put, and, more significantly, the basic framework for social science.

Finally, the census is both one of the oldest survey formats for social science information, and in a remarkable number of cases, the historical data, including original questionnaires, records of administrative procedures, as well as the publicly released results, have survived and are available for current use. The availability, therefore, of repeated collections of what were originally cross-sectional data collections permits sophisticated longitudinal analysis of the information in the censuses.

4 The census as a source for longitudinal analysis

Censuses, as official and public sources of information, serve many masters. Chief among those masters are the governmental or official interests that fund the data collection and make initial use of the results. The official grounding of census data collections in turn guarantees that in nations committed to collecting and disseminating census data, researchers can depend upon the quality, public availability, and consistency of the data. For example, in the United States for over two centuries, researchers have been able to count upon the release of tabulated population data from the census. Since the 1960s, they have been able to count upon the availability of public use micro data samples from the complete count (PUMS files) for additional research use. Interestingly, however, for the researcher intent upon using census information for longitudinal analysis, the very strengths of the data source, e.g., accessibility, comprehensiveness, and long temporal run of data, also necessarily present significant methodological issues that must be tackled before one can make effective use of the data. In other words, the strengths of this data source are inextricably tied to its weaknesses, complications, and frustrations. Thus a significant methodological literature exists to guide the researcher through the minefield of methodological problems and facilitate analysis, and it is to these issues that we now turn.

5 Data availability

Longitudinal analysis requires repeated collection of consistent data over time. For a researcher intending to use census data, his or her first task is to determine when the censuses were taken and the amount and type of data preserved and available for current use. Population censuses are fundamentally an activity of modern states in the West, and date primarily from 1800 and later. Researchers interested in longitudinal analysis of population and economic relationships before 1800 must identify additional sources with a longer collection history. These include, for example, parish registers, occasional intermittent censuses, or administrative data, e.g., records of tax collections, or the heights, weights, and personal information collected of military recruits.

By the middle of the twentieth century, governments around the world instituted periodic censuses and the United Nations has summarized this information in its annual publication, the *Demographic Yearbook*, and maintains links to national statistics offices on its website. The United States Census Bureau website also has links to the national statistical offices of most nations for the researcher interested in finding census data on a particular country or region.

Once a researcher has determined that the historical census data were collected for the time period of interest, he or she will face an additional set of technical challenges. The first question of importance is whether the longitudinal data file to be created is made up of aggregate or micro level data.

6 Aggregate and microdata

6.1 Aggregate

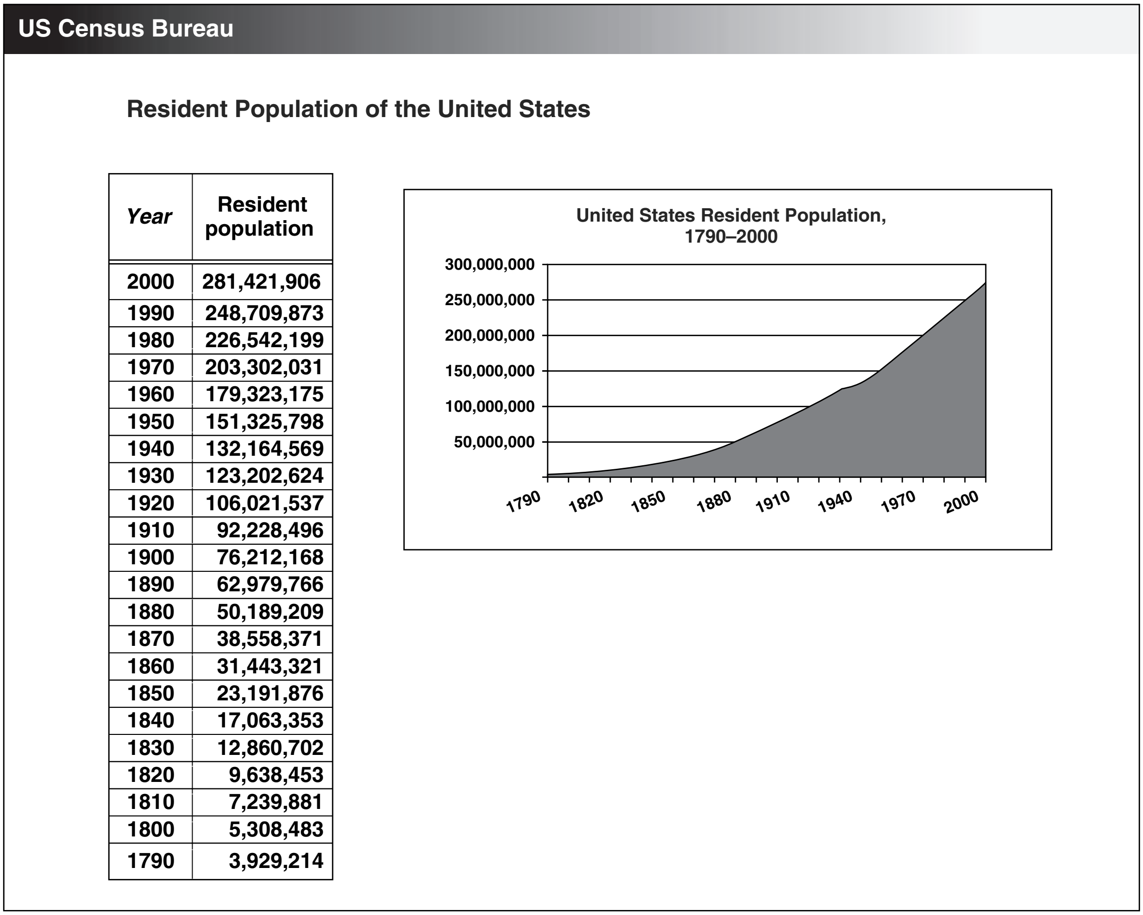
The official agencies that collect the individual level census information generally publish tabulations of the results in the years immediately after the count. These data, when aggregated for multiple censuses across time, can provide the aggregate time series of basic trends for the geographic area, industry, or population subgroup of interest. Researchers must then determine if the reporting categories were consistent for a number of points in time. At the simplest level, one can prepare a time series of aggregate population for a country, or for basic reporting categories, e.g., for the number of males or females in a nation's population. National statistical agencies tend to publish such trend data in the main or supplementary publications of the census or in retrospective compilations. See Figures 2.1 and 2.2, which report time series results in tabular and graphic form for the total United States population and for several basic demographic variables. (Additional detail is available at "Selected Historical Decennial Census Population and Housing Counts," http://www.census.gov/population/www/censusdata/hiscendata.html.)

Retrospective compilations provide more elaborate time series, and also include technical analyses explaining the compilation of individual series. These compilations require substantial effort to revisit the quality of individual data points in the series, explain breaks and ambiguities, and trace the development of classifications and concepts. See, for example, *Historical Statistics of the United States* (HSUS), US Bureau of the Census 1976. The most recent revision of the US compilation is an ambitious scholarly collaboration (Susan Carter, et al., 2006).

For more complex analyses, reporting categories must remain consistent from census to census. The researcher may have to prepare the time series by retrieving each data point in the time series from each year's published census volume or table. For example, if a researcher is interested in aggregate population change for cities for a nation, he or she will face the issues that the number of cities tends to grow over time and that the geographic boundaries of cities also change.

To use a simple example, the US census has reported the population of New York City since 1790. The boundaries of the modern city of five boroughs (Manhattan, Brooklyn, Queens, Bronx, Richmond (Staten Island)) were defined in 1898. Thus the researcher must decide whether to include in the series comparative data for the period from 1790 to 1890, or whether to report only the modern boundaries. See Table 2.1 and Figures 2.3 and 2.4 which illustrate the issues. Before the 1900 census, New York City included New York County (Manhattan) and a small portion of Westchester County (a portion of what is now the borough of the Bronx), which was annexed to the city in the 1870s. If the researcher uses the modern boundaries, the city's population shows a huge jump between 1890 and 1900, primarily because the creation of Greater New York annexed the nation's fourth largest city (Brooklyn) to New York City. The remaining annexed areas of Staten Island, Queens, and the Bronx grew rapidly in later years. More complex aggregation decisions occur when the geographic boundary changes are more complex, or when the researcher wishes to aggregate on demographic or economic variables, as discussed below.

In short, researchers interested in preparing aggregate longitudinal data series are dependent upon the reporting and publication decisions of the national statistical agencies that compiled the data. If the original data were not tabulated and reported on a researcher's category of interest, the data series cannot be created from the published results. In the face of this problem, many researchers have turned to accessing micro level data from past censuses with the goal of aggregating the results into categories of interest, or using the microdata directly for analysis. These data in turn present different challenges for longitudinal analysis.



US Census Bureau

Resident Population of the United States

Year	Resident population
2000	281,421,906
1990	248,709,873
1980	226,542,199
1970	203,302,031
1960	179,323,175
1950	151,325,798
1940	132,164,569
1930	123,202,624
1920	106,021,537
1910	92,228,496
1900	76,212,168
1890	62,979,766
1880	50,189,209
1870	38,558,371
1860	31,443,321
1850	23,191,876
1840	17,063,353
1830	12,860,702
1820	9,638,453
1810	7,239,881
1800	5,308,483
1790	3,929,214

Figure 2.1 United States Census Bureau population growth with graph, 1790 to 2000, http://www.census.gov/dmd/www/resapport/states/unitedstates.xls

6.2 Microdata

Longitudinal analysis of microdata census results are possible only if the collecting agency has preserved the individual responses either on the original paper schedules, on microfilm or related media, or in electronic form. The researcher also needs to be able to gain access to these data for research use. That is, the country under study must provide a mechanism to make the data available. In general, national statistical agencies have confidentiality and privacy protections which prevent non-official access to the individual level responses to the census questions, e.g., to the microdata. These strictures have been in place since the early twentieth century to prevent abuse of the individual respondent's privacy rights. Though rules for research access to microdata from past censuses vary by country, many countries will preserve confidentiality for a period of years, and then allow researcher and public access. For the United States, for example, the individual level information on the population census schedules is protected from public access for 72 years after the census is taken. Thus in the United States, researchers can access microfilm

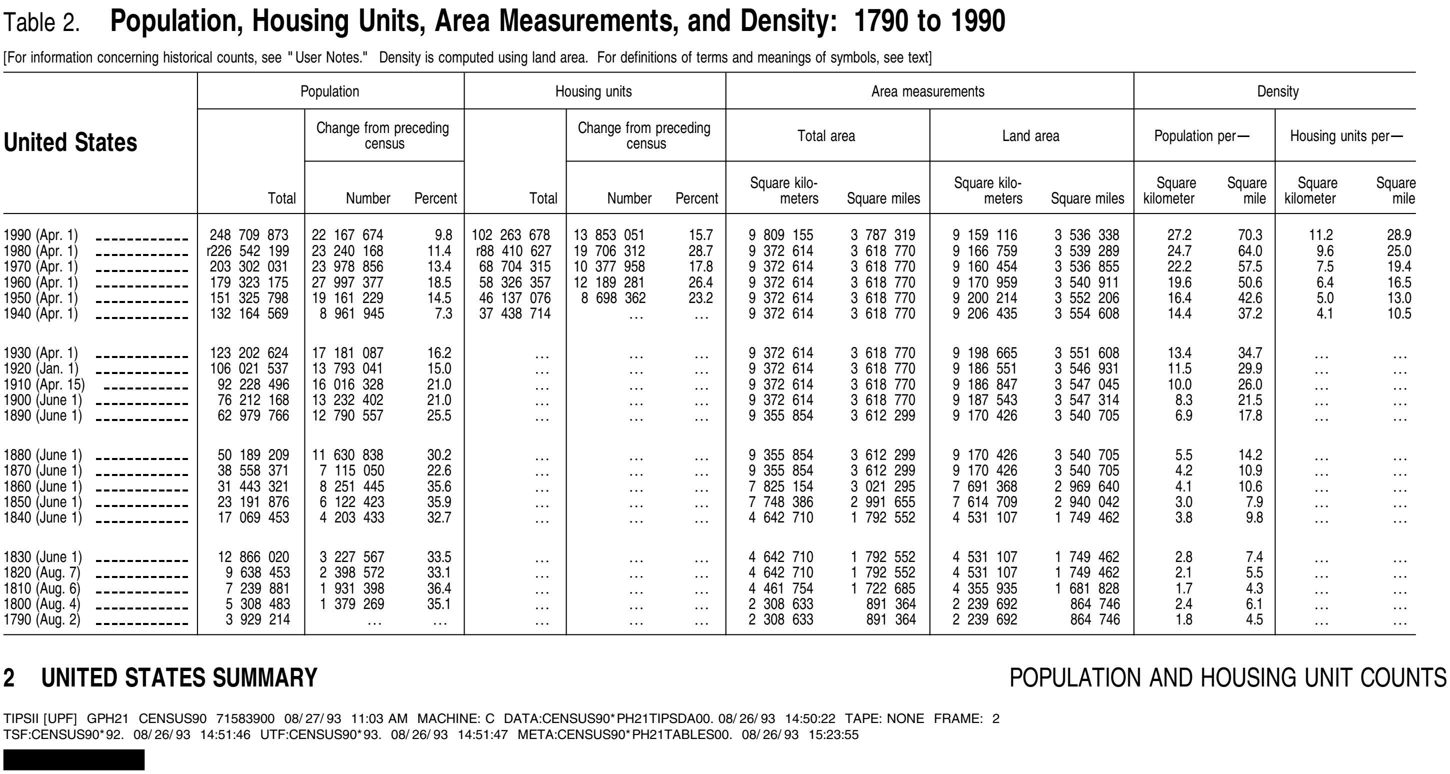


Table 2. **Population, Housing Units, Area Measurements, and Density: 1790 to 1990**

[For information concerning historical counts, see "User Notes." Density is computed using land area. For definitions of terms and meanings of symbols, see text]

United States	Population			Housing units			Area measurements				Density			
		Change from preceding census			Change from preceding census		Total area		Land area		Population per—		Housing units per—	
	Total	Number	Percent	Total	Number	Percent	Square kilo-meters	Square miles	Square kilo-meters	Square miles	Square kilometer	Square mile	Square kilometer	Square mile
1990 (Apr. 1)	248 709 873	22 167 674	9.8	102 263 678	13 853 051	15.7	9 809 155	3 787 319	9 159 116	3 536 338	27.2	70.3	11.2	28.9
1980 (Apr. 1)	r226 542 199	23 240 168	11.4	r88 410 627	19 706 312	28.7	9 372 614	3 618 770	9 166 759	3 539 289	24.7	64.0	9.6	25.0
1970 (Apr. 1)	203 302 031	23 978 856	13.4	68 704 315	10 377 958	17.8	9 372 614	3 618 770	9 160 454	3 536 855	22.2	57.5	7.5	19.4
1960 (Apr. 1)	179 323 175	27 997 377	18.5	58 326 357	12 189 281	26.4	9 372 614	3 618 770	9 170 959	3 540 911	19.6	50.6	6.4	16.5
1950 (Apr. 1)	151 325 798	19 161 229	14.5	46 137 076	8 698 362	23.2	9 372 614	3 618 770	9 200 214	3 552 206	16.4	42.6	5.0	13.0
1940 (Apr. 1)	132 164 569	8 961 945	7.3	37 438 714	...	...	9 372 614	3 618 770	9 206 435	3 554 608	14.4	37.2	4.1	10.5
1930 (Apr. 1)	123 202 624	17 181 087	16.2	...	...	...	9 372 614	3 618 770	9 198 665	3 551 608	13.4	34.7	...	...
1920 (Jan. 1)	106 021 537	13 793 041	15.0	...	...	...	9 372 614	3 618 770	9 186 551	3 546 931	11.5	29.9	...	...
1910 (Apr. 15)	92 228 496	16 016 328	21.0	...	...	...	9 372 614	3 618 770	9 186 847	3 547 045	10.0	26.0	...	...
1900 (June 1)	76 212 168	13 232 402	21.0	...	...	...	9 372 614	3 618 770	9 187 543	3 547 314	8.3	21.5	...	...
1890 (June 1)	62 979 766	12 790 557	25.5	...	...	...	9 355 854	3 612 299	9 170 426	3 540 705	6.9	17.8	...	...
1880 (June 1)	50 189 209	11 630 838	30.2	...	...	...	9 355 854	3 612 299	9 170 426	3 540 705	5.5	14.2	...	...
1870 (June 1)	38 558 371	7 115 050	22.6	...	...	...	9 355 854	3 612 299	9 170 426	3 540 705	4.2	10.9	...	...
1860 (June 1)	31 443 321	8 251 445	35.6	...	...	...	7 825 154	3 021 295	7 691 368	2 969 640	4.1	10.6	...	...
1850 (June 1)	23 191 876	6 122 423	35.9	...	...	...	7 748 386	2 991 655	7 614 709	2 940 042	3.0	7.9	...	...
1840 (June 1)	17 069 453	4 203 433	32.7	...	...	...	4 642 710	1 792 552	4 531 107	1 749 462	3.8	9.8	...	...
1830 (June 1)	12 866 020	3 227 567	33.5	...	...	...	4 642 710	1 792 552	4 531 107	1 749 462	2.8	7.4	...	...
1820 (Aug. 7)	9 638 453	2 398 572	33.1	...	...	...	4 642 710	1 792 552	4 531 107	1 749 462	2.1	5.5	...	...
1810 (Aug. 6)	7 239 881	1 931 398	36.4	...	...	...	4 461 754	1 722 685	4 355 935	1 681 828	1.7	4.3	...	...
1800 (Aug. 4)	5 308 483	1 379 269	35.1	...	...	...	2 308 633	891 364	2 239 692	864 746	2.4	6.1	...	...
1790 (Aug. 2)	3 929 214	...	...	...	...	...	2 308 633	891 364	2 239 692	864 746	1.8	4.5	...	...

2 UNITED STATES SUMMARY POPULATION AND HOUSING UNIT COUNTS

TIPSII [UPF] GPH21 CENSUS90 71583900 08/27/93 11:03 AM MACHINE: C DATA:CENSUS90*PH21TIPSDA00. 08/26/93 14:50:22 TAPE: NONE FRAME: 2
TSF:CENSUS90*92. 08/26/93 14:51:46 UTF:CENSUS90*93. 08/26/93 14:51:47 META:CENSUS90*PH21TABLES00. 08/26/93 15:23:55

[redacted]

Figure 2.2 United States Census Bureau historical population tables, http://www.census.gov/dmd/www/resapport/states/unitedstates.xls

Table 2.1 New York City population, 1790–2000, breakdown by boroughs

	Year	*Total*	*Bronx*	*Brooklyn*	*Manhattan*	*Queens*	*Staten Is*	*Old NYC*
1	1790	33131	–	4495	–	–	3835	33131
2	1800	60489	–	5740	–	–	4564	60489
3	1810	96373	–	8303	–	–	5347	96373
4	1820	123706	–	11187	–	–	6135	123706
5	1830	202589	–	20535	–	–	7082	202589
6	1840	312710	–	47613	–	–	10965	312710
7	1850	515547	–	138882	–	–	15061	515547
8	1860	813669	–	279122	–	30249	25492	813669
9	1870	942292	–	419921	–	41669	–	942292
10	1880	1206299	–	599495	–	52927	–	1206299
11	1890	2507414	88908	838547	1441216	87050	51693	1515301
12	1900	3437202	200507	1166582	1850093	152999	67021	1850093
13	1910	4766883	430980	1634351	2331542	284041	85969	2331542
14	1920	5620048	732016	2018356	2284103	469042	116531	2284103
15	1930	6930446	1265258	2560401	1867312	1079129	158346	1867312
16	1940	7454995	1394711	2698285	1889924	1297634	174441	1889924
17	1950	7891957	1451277	2738175	1960101	1550849	191555	1960101
18	1960	7781984	1424815	2627319	1698281	1809578	221991	1698281
19	1970	7894862	1471701	2602012	1539233	1986473	295443	1539233
20	1980	7071639	1168972	2230936	1428285	1891325	352121	1428285
21	1990	7322564	1203789	2300664	1487536	1951598	378977	1487536
22	2000	8008278	1332650	2465326	1537195	2229379	443728	1537195

Source: New York City Department of City Planning, Change in Total Population, 1990 and 2000, New York City and Boroughs, http://www.ci.nyc.ny.us/html/dcp/html/census/pop2000.shtml. Ira Rosenwaike, *Population History of New York City* (Syracuse, NY: Syracuse University Press, 1972); United States Census Bureau, 'Population of the 100 Largest Cities and Other Urban Places in the United States: 1790 To 1990', http://www.census.gov/population/www/documentation/twps0027.html

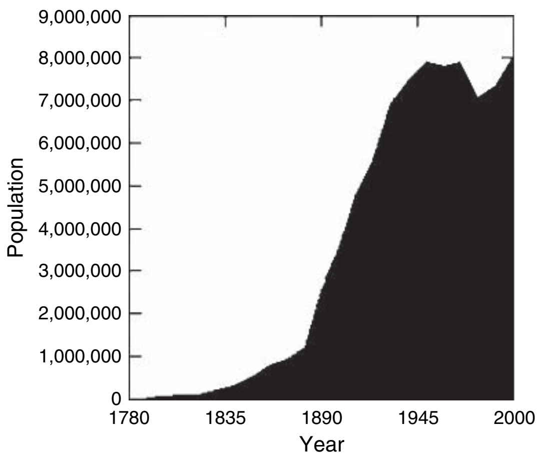


Figure 2.3 New York City population, 1790–2000

copies of the population schedules for the censuses from 1790 to 1930 (with the exception of the schedules for the 1890 census which were burned in a fire in the early 1920s). The Canadians release data after 92 years.

Starting with the 1960 census, the US Census Bureau also prepared a public use file, a 1% or 5% sample of the complete count. Data in the samples are coded to protect confidentiality such that no individual's information can be identified in the file prior to the 72-year limit. To prevent disclosure, the agency restricts the geographic detail available on the cases of the sample.

Since the 1970s, with grant funding chiefly from the National Institutes of Health and the National Science Foundation, a series of

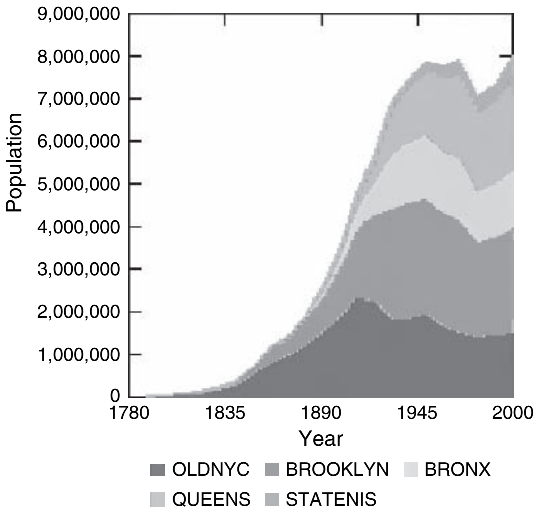


Figure 2.4 New York City population, 1790–2000, breakdown by boroughs

researchers, most notably the demographic historians at the Minnesota Population Center at the University of Minnesota, have created public use files of past US censuses for 1850–1880, and 1900–1950, and have developed standardized coding and web-based file delivery systems to create "integrated public use microdata samples" for the US population censuses from 1850 to 2000. These data, with documentation, bibliographies of research use, facsimiles of original questionnaires, and technical papers, are available at the IPUMS website, http://www.ipums.org. Other nations have also prepared public use files of their censuses. For a current inventory of population censuses with information on the current knowledge about surviving microdata files, see http://www.ipums.org/international/microdata_inventory.html. The Minnesota Population Center has also developed an "international integrated public use microdata sample" project or IPUMS International, http://www.ipums.org/international/ index.html to develop a web-based harmonized data delivery system similar to that developed for the United States population censuses. These microdata files almost all date from 1960 or later and thus include only data points from four to five censuses. Only a few nations (Table 2.2) have extant microdata files from before 1960. The United States has the most with 15. Norway has the longest data run, with the early census data from 1801. One can expect that in the years to come additional files will be created, as the promise of developing such longitudinal series has proved itself in the past 15 years.

Table 2.2 Availability of data points in historical PUMS files

Country	*Data points*	*Span of data*
Norway	12	1801–2001
United Kingdom	7	1851–2001
Canada	12	1871–2001
Argentina	7	1869–2001
Finland	9	1950–2001
United States	15	1850–2000

Source: Integrated Public Use Microdata Series International, Census Years and Microdata Inventory, April 1, 2006, http://www.ipums.org/international/microdata_inventory.html

7 Questions

7.1 Availability

A second methodological issue facing researchers using longitudinal census data is the availability and character of the questions asked and reported over time. Broadly speaking, the number of questions and the amount of detail collected increases over time so that more information is available in current census reports than in older ones. For example, the first census of the United States (1790) asked only for the name of the household head and for five additional pieces of information for the household (the number of free white males 16 and over; the number of free white males under 16; the number of free white females; the number of slaves, and the number of other free people). By 1850, each person in the United States was identified on a separate census line. The census asked 13 questions of the free population and 8 for the slave population. The resulting data provided detailed information on the address, household relationships, age, sex, race, nativity, occupation, educational, marital and disability status, and property owned. See http://www.ipums.org/usa/voliii/items1850.html for the 1850 schedule for the free population and http://www.ipums.org/usa/voliii/tEnum Form.shtml for access to all the forms from 1850 to 2000. For the facsimiles of the slave schedules, and for pre-1850 census schedules, see US Bureau of the Census, 1979; US Bureau of the Census, 1973. See also United States Census Bureau, "Selected Historical Decennial Census Population and Housing Counts," http://www.census.gov/population/www/censusdata/hiscendata.html, for links to forms and enumerator instructions from 1790 to 2000.

In 2000, respondents to the US census answered six basic questions: name and address, household relationship, age and place of birth, sex, race and ethnicity, and housing tenure (whether the dwelling was owned or rented). A one in six sample of households received a "long form" questionnaire with a total of 53 questions on demographics, housing conditions, occupation and income, migration, citizenship and languages spoken, educational attainment, disability status. For copies of the forms, see http://www.census.gov/dmd/www/2000quest.html.

The IPUMS website, www.ipums.org, and *Twenty Censuses* and *Population and Housing Inquiries in U.S. Decennial Censuses, 1790–1970* (US Bureau of the Census, 1979; 1973) provide tables listing the availability of particular American census questions over time. The IPUMS site also lists coding schemes for answers over time. Thus a researcher can determine if data were collected consistently on a particular item of interest, and how the question was framed and responses recorded at each census.

Some basic questions, e.g., age and sex, are relatively straightforward and thus the responses can be assumed to be historically consistent. The basic questions and the coding schemes are fundamentally consistent over time though even with such basic questions, methodological problems, e.g., age heaping, can exist in the original data. Other questions, even such seemingly transparent questions such as race, place of birth or educational attainment, are not, on closer inspection. Three types of issues emerge: conceptual change in the phenomenon measured; changes in classification schemes; and level of detail in the answers. A number of examples illustrate the issues.

7.2 Changes in conceptualization of a phenomenon

Longitudinal researchers are often interested in the origins of a phenomenon, asking when something started, or, relatedly, when it ended. For census questions, the measurements of economic and educational status illustrate the issues. Mass public education is an innovation of the nineteenth century in modern societies, and thus for some older censuses, the earliest questions on the educational characteristics of the population asked a yes/no question about literacy, e.g., whether someone could read or write, or was literate in a particular language. By the twentieth century, the question changed to one of educational attainment, e.g., the number of years of schooling for an individual.

Similar changes occurred in questions on economic status. The United States, for example, asked questions about the ownership of real and personal property from 1850 to 1870 and income from 1940 and later. The main question on economic status for the majority of historical censuses, however, was a question on occupation, first asked in 1820, and reported consistently since 1840. For the censuses from 1850 on, the individual level occupation question was open ended. For such questions there were also frequently age and sex thresholds, which limited the question to adults, males, household heads, or individuals capable of owning property.

Such changes in questions over time present difficulties for longitudinal analysis and require the researcher to evaluate the meaning of the changes. Different strategies are appropriate depending on whether the variable is the central focus of the analysis or a correlate or control for the analysis of another item of interest.

For analyses where the object of study is changing over time, the shifting variables are themselves evidence of the historical phenomenon of interest. Asking a mid-nineteenth-century American about his income would have elicited a confused response, since most people did not work for a wage or salary. Similarly, asking for the number of years of schooling completed made little sense when schooling was intermittent and schools scarce.

For researchers interested in using a measure of economic or social status as a correlate for changes over time, a number of techniques exist to develop a consistent measure. Most common is the scale measuring socioeconomic status, SES or SEI, which assigns a numeric value to an occupational code, and thus can be used to assign a common variable code for economic status for census results in which occupation has been recorded. See the discussion of the SEI variable in the IPUMS data, http://www.ipums.org/usa–action/variableDescription.do?mnemonic-SEI.

7.3 Classification changes

A related issue for longitudinal analysis of census responses involves changes in classification and coding schemes over time. Similar to the changes in the conceptualization of the question, classification changes may represent the origin or elimination of a phenomenon. The most complex coding scheme again involves the occupation variable. Computer programmers did not exist in the nineteenth century, nor do "intelligence office keepers" exist today.[1] Thus the researcher requires a standardized coding scheme to accommodate the change represented in the occupation codes themselves. The IPUMS project has confronted these issues and has developed several systems for longitudinal occupation coding. Additional guidance on the methodological issues involved is available on their website, http://www.ipums.org.

Similar problems plague coding schemes for geographic variables, e.g., place of birth or current address. As with the discussion above of the changes in city boundaries, the researcher must decide how to code and evaluate addresses and geographic codes which change over time. For example, for a study of occupational and geographic mobility in the United States, a person born in 1905 in the Russian empire might have emigrated to the US from the Soviet Union of 1921, from a town which today is in Ukraine.

Even variables with more limited coding schemes, such as race, are significantly more complex when used as longitudinal variables. Margo Anderson and Stephen E. Fienberg (2001, pp. 177–78) list the race categories from the US census from 1820 to the present. A person in this scheme could change his or her race from census to census because of changes in the coding conventions.

7.4 Detail

A final issue of comparability over time in longitudinal census data involves the level of detail reported on particular variables. Older censuses from the pre-computer age tended to employ simple response categories for a closed end question and employ few follow-up questions to clarify a phenomenon. For questions on disability, for example, older censuses asked if a person was blind, deaf, "insane or idiotic."

[1] "Intelligence office keepers" ran employment agencies (Margo Anderson Conk, 1980).

Current questions on disability include detailed information on the nature of the disability, and its impact in terms of functional impairment in the activities of daily living, and the duration of disability.

Thus, though it is often possible to get a fairly lengthy and detailed time series of data by aggregating data from multiple censuses, the researcher must expect to put in significant effort in understanding any changes in question concept, coding, and reporting over time, and reconcile the differences among the various schemes before analysis. Not surprisingly, therefore, substantial methodological literature exists on solving these problems. In particular areas, for e.g., historical race classification, this literature has formed a substantive subfield in its own right as part of the history of race in modern societies. See, for example, Joel Perlmann and Mary Waters, 2002; Melissa Nobles, 2000; Clara Rodriguez, 2000; F. James Davis, 1991; Anderson and Fienberg, 2001; Tukufu Zuberi, 2001.

8 Uses of census data in research

8.1 Trends

The most common form of research using longitudinal census data is the time series or analysis of trends over time. The simplest of such analyses chart the changes over time of reported data either in tabular or graphic form with calculations of percentage or absolute change between data points (Figure 2.1). Such reports of aggregate trends, particularly when graphed, almost naturally raise questions of the determinants, correlates or causes of the changes. The classic linear regression model became the fundamental technique to explore such temporal change. Since most census data record rapid population growth in the nineteenth and twentieth centuries, such analysis of aggregates has focused on analyzing the rates and volume of change, whether stable or erratic. Such analysis makes it possible to identify the impact of historical events external to the census on the data series. War, economic depression, or changes in a nation's population policy on immigration, can then be identified in the series, and the relevant variables included in the model to test various hypotheses.

See, for example, Walter Nugent (1981) for the analysis of the growth patterns displayed in the decennial census. In the basic graph of population change for the US, the United Kingdom and France from 1790 to 2000 (Figure 2.5), the dramatic numeric increase in the US population is the dominant lesson of the image. Regraphing on the rate of change by census year reveals a different pattern (Figure 2.6), namely two periods of stable rates of growth connected by a declining trend between the two. Nugent used the patterns to explore the transition of the United States from an agricultural and "frontier rural" nation to an urban nation. Roderick Floud (1979, pp. 88–137) provides a related example in an analysis of British domestic exports from 1820–1850.

Trend patterns can also be displayed geographically. The United States Census Bureau has, since the 1870s, reported a statistic on the "center of population" for the nation.

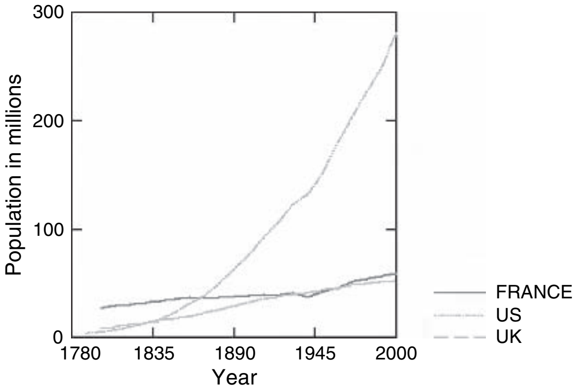


Figure 2.5 Population of the UK, US and France, 1790–Present

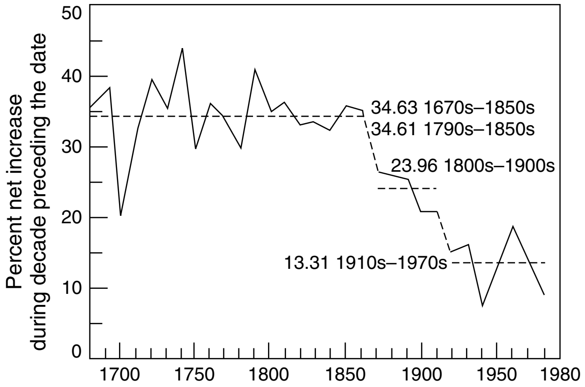


Figure 2.6 Decomposing growth trends, United States, 1790–2000 (reproduced with permission)
Source: Walter Nugent (1981). *Structures of American Social History*. Bloomington: Indiana University Press http://images.questia.com/fif=b48586/b48586p0027.fpx&init=0.0,0.0,1.0,1&rect=0.0,0.25,0.5,0.75&wid=300&hei=205&vtrx=1&lng=en_US&enablePastMaxZoom=OFF&xFactor=100&page=qView,html&obj=uv,1.0&cmd=ZOOM_OUT

Theoretically, the "center of population," is the "place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly" if everyone were of identical weight. http://www.census.gov/geo/www/cenpop/cb01cn66.html. See Figure 2.7, which graphs the mean center of population for each census since 1790. The starred data point moved westward from 1790 to 1880, then slowed as the frontier closed and the country urbanized until 1940, then moved southwestward into Missouri by 2000 with the growth of the Sunbelt. The use of a geographic visualization of a temporal pattern of change provides a powerful analysis of longitudinal data.

8.2 Denominator data

Longitudinal census data are also frequently used as denominators in the analysis of other issues of interest, e.g., voting patterns, vital rates, or disease patterns. In such cases, the object of study requires data on the total population from which the population of interest is drawn. For example, the analysis of electoral behavior and voter participation requires information on the overall population from which actual voters are drawn. Historians in the United States have made good use of local area census results from the nineteenth and twentieth centuries to chart and analyze the growth of mass democracy and electoral turnout, to identify the determinants of voter mobilization and party success in particular local areas, and to develop the theory of critical elections and electoral realignment. For detail on the substantive results of this literature, see Theodore Rosenoff (2003).

The need for systematic development of longitudinal datasets for electoral analysis was recognized in the 1960s. Political scientists who recognized the need for such datasets founded the Inter-University Consortium for Political and Social Research 1962 to coordinate the creation and preservation of electronic social science data files which were too large for any single researcher to assemble. Its first compilations were local area US presidential election returns from 1824 on, and candidate and constituency totals from 1788 on (ICPSR 1 and ICPSR 2). The third data file in the archive (ICPSR 3) compiled the US census denominator data required for electoral analysis, i.e., "detailed county and state-level ecological or descriptive data for the United States" from the published decennial census volumes from 1790 forward.[2]

[2] Inter-university Consortium for Political and Social Research, United States Historical Election Returns, 1824–1968 (ICPSR 1); Candidate Name and Constituency Totals, 1788–1990 (ICPSR 2); Historical, Demographic, Economic, and Social Data: the United States, 1790–1970 (ICPSR 3) [Computer files]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], last updated 1999, 1995, 1992 respectively. (See www.icpsr.umich.edu.)

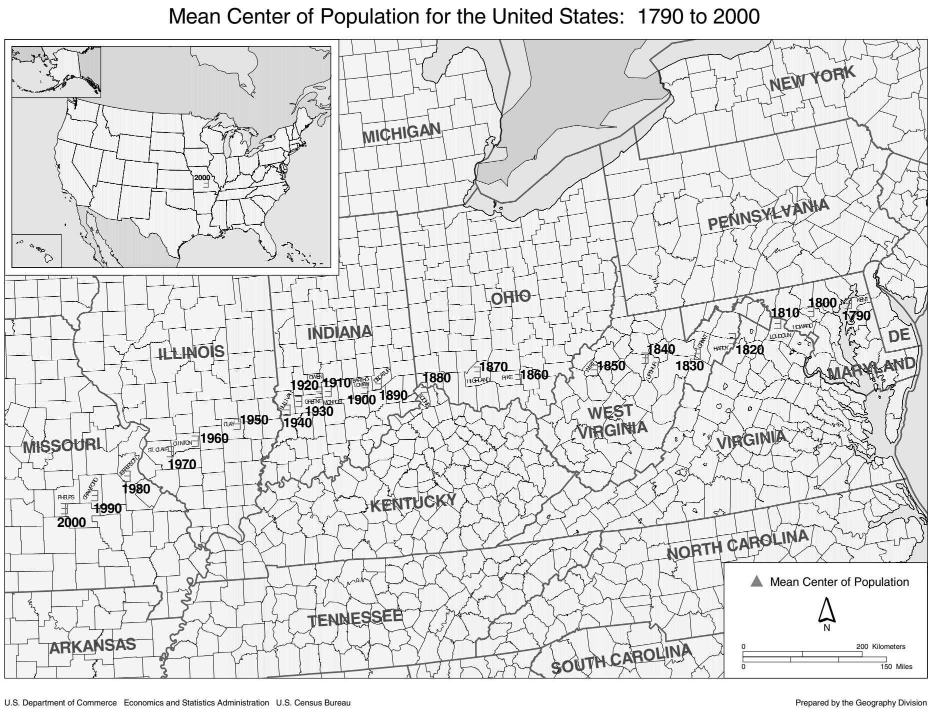


Figure 2.7 United States center of population, 1790–2000, http://www.census.gov/geo/www/cenpop/cbøkn66.html, http://www.census.gov/geo/www/cenpop/meanctr.pdf.

8.3 Record linkage

Theoretically, censuses count the entire population of a nation at regular intervals and record the name and address of individuals. For censuses no longer covered by confidentiality protections, it should be possible to locate particular individuals in multiple censuses, or with other contemporary data sources such as city directories, property records, or related data. Genealogists, of course, pursue such linkages in search of family lineages, and family historians have also made good use of such records to construct family histories. Substantial methodological resources have been developed to facilitate such research. For example, it is possible to search US census manuscripts by the name of the person using the SOUNDEX indexes.

Other historians have conducted such research with mixed results because of the weaknesses in using census data for such linkage. The problems are several. First, the research is extremely time-consuming. Though copies of the manuscript schedules are available in research libraries, individuals need to be traced through multiple originally handwritten data sources on microfilm. Second, matching individual cases can be problematic because of name spellings and name changes in different censuses. These are particular problems for searching for women who change name at marriage. Common names, e.g., John Smith, require additional information from related variables, e.g., an age or occupation link, to confirm a match. And, as many scholars discovered, census coverage is incomplete, and estimates of the level of incompleteness are quite imprecise before the 1940s.

The result of a substantial number of studies which attempted to link individuals over time and trace occupational mobility from fathers to sons was the realization that one of the major findings of the research was the evidence of the volatility of the population. Some 40–80% of the cases could not be traced, and thus confirmed a different picture of historical mobility from the one initially hypothesized. See, for example, Stephan Thernstrom, 1964; Avery Guest, 1987; Peter Knights, 1991a; 1991b; Joseph Ferrie, 1996.

Currently underway are new projects based upon more advanced methods to attempt to improve linkages between census dates. For further details, see Steven Ruggles (2003a). In the UK, the Office of National Statistics has built a prospective individual level longitudinal linked sample database (the Longitudinal Study or LS) from the decennial censuses of 1971 and later. There are confidentiality restrictions on these data (see the discussion above on confidentiality issues in longitudinal census research). See http://www.celsius.lshtm.ac.uk/what.html for details on the files and their use.

8.4 Sampling frames

Since the development of probability sampling methods in the 1930s, the results of the complete count census have been used to inform the frames for sample surveys, both officially within governments and by private researchers. To serve this function, the accuracy of the complete count data must be assured. That is, the researchers creating the sample need estimates of overcount, undercount, and bias in the underlying census data collection. Modern central statistical agencies provide such estimates. Researchers interested in drawing samples from data collected before the 1930s, either in the census itself or from related historical data such as property records or vital records, will generally have to decide how to approach the question of census accuracy for their particular design. For example, for a discussion of this issue, in the context of the analysis of fertility decline in the United States, see Hacker (2003).

9 The potential of longitudinal analysis of census data

The literature cited in the discussions above provides evidence of the importance that scholars and officials within central statistical agencies have placed on the longitudinal analysis of census data. The availability in the last 10 to 15 years of much more extensive electronic data files, and the potential of many more to come, has boosted the interest in the field, and is beginning to produce important results. Illustrative of this work is James Gregory's recent book, *The Southern Diaspora* (2005), which expands the understanding of migration out of the American South in the twentieth century.

Studies of the large-scale Black migration from the agricultural areas of the American South to the cities of the North and West are well known, as are numerous case studies of migration, such as of whites from Appalachia to the industrial Midwest or the Okies to California. But until the availability of IPUMS microdata, it has not been possible to trace systematically the massive out migration from the American South. Twenty-eight-million people (Gregory, 2005, pp. 19, 330) left the American South for the North and West in the twentieth century, dramatically changing the economy, culture, and politics of the region they left and the regions they settled. The newly available longitudinal microdata make it possible to identify the occupational and educational characteristics of the migrants, as well as explore family structure, return migration, and the impact of the migrants on the receiving communities. Among his findings are that the white southern diaspora is considerably larger than the black diaspora, but less visible since southern whites tended to migrate to smaller cities and be less visible migrants in their new destinations.

Patricia Hall and Steve Ruggles (2004) have produced equally provocative results from their analysis of longitudinal IPUMS data. They return to Frederick Jackson Turner's famous analysis of the significance of the frontier in American history, exploring internal migration from 1850 to 2000, confirming much of Turner's argument, while noting the impact of a new wave of suburban migration in the second half of the twentieth century. See also Steven Ruggles (2003b) on longitudinal analysis of US family structure.

10 Historical context and the politics of numbers

As noted at the outset, census data are collected by the national state and thus have the authority of the state and the interests of the state embedded in them. In all states, whether democratic, authoritarian, or imperial, a politics of numbers frames the data collection, reporting, and preservation process, which in turn affects potential historical research from the data years and centuries later. A researcher can trace the contemporary political controversies of the day in the methods and reports published at the time of the census, and is well advised to spend some time understanding the issues in order to judge the quality and capacities of the data for longitudinal analysis. One might expect that the censuses from culturally similar, primarily English-speaking societies of Britain, Canada and the US, would be quite similar, and in many ways they are. These nations all have long traditions of data collection, quite similar enumeration methods, and practices of publishing extensive census results. Nevertheless, there are important differences in emphasis, questions, and reporting styles for the data, which derive from the different uses and political traditions of their national states.

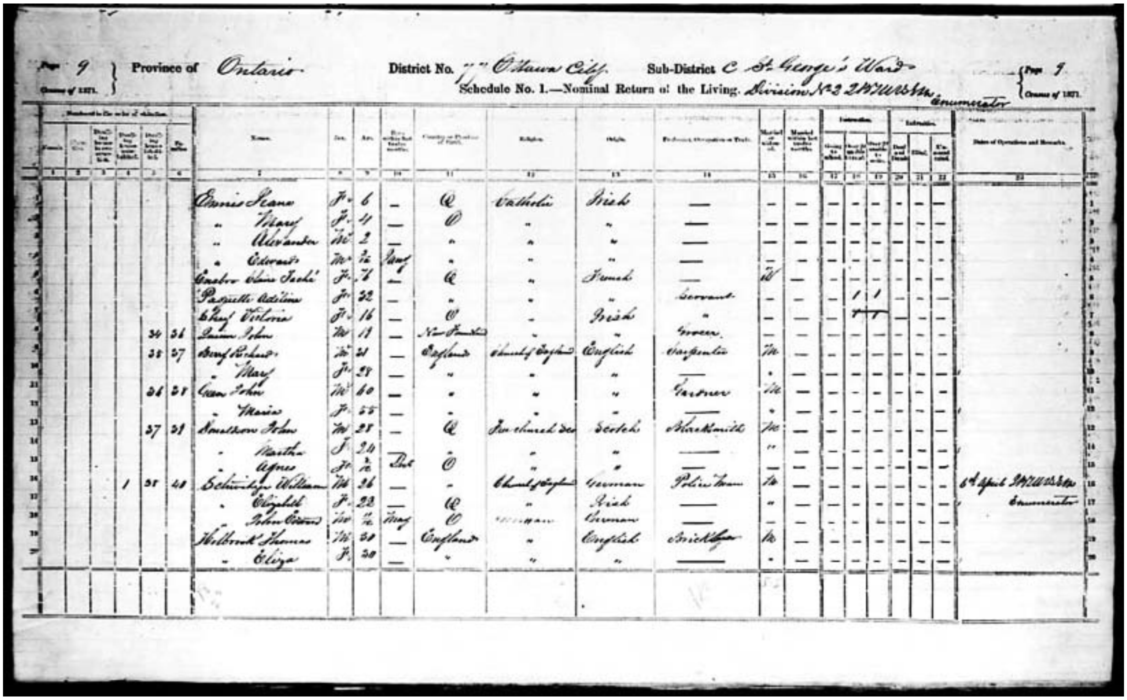
The American census is the oldest compilation and the first census to be used for the purposes of legislative apportionment. Accordingly, the questions asked and the tabulations prepared have always been closely related to the political and demographic controversies of the society. In the nineteenth century, these were issues of race, population growth and change, and migration (Margo Anderson, 1988; Anderson and Fienberg, 2001). The US census has recorded the racial characteristics of the total population directly since 1820, and of the white population from 1790 to 1810. At the height of the early twentieth century immigration surge, the census asked 10 questions on immigrant status, mother tongue and English-speaking ability, and citizenship status. The United Kingdom census, by comparison, did not ask a question on the "ethnic group" of the respondent until 1991.

By contrast, the situation and evolution of social classes underpinned much of the questions and analysis in the UK census in the nineteenth and early twentieth centuries (Edward Higgs, 1996; Simon Szreter, 1996). In the first urban nation, and the home of the industrial revolution, precise information on respondents' occupations and occupational classifications were developed to provide denominator data to evaluate the fertility, mortality and overall health of the working classes and middle classes.

In the Canadian census one sees a third interest, namely in the language background of a population that was created when English and French colonies joined to form a national state (Bruce Curtis 2001). At the census of Canada in 1871 (Figures 2.8 and 2.9, http://www.collectionscanada.ca/genealogy/022-911-e.html,) questionnaires were prepared in French and English. In 2001, the Canadian census had questions on the first language learned in childhood, languages understood and spoken at home, knowledge of official and non-official languages in the various regions of Canada, and on language used at work.

A related issue in the use of census data for longitudinal analysis involves recognizing that the published data may be tainted in some form. For example, it is generally recognized that the reports of insanity by race for the US census of 1840 are wrong and there have been major controversies about the quality of the population count in the American South at the 1870 census, so much so that the Census Bureau "corrected" the totals in later reprints. For details, see Patricia Cline Cohen, 1982; Anderson, 1988; Hacker, 2003.

In other words, if a researcher intends to create his or her own longitudinal data file by compiling reported data from past censuses, or by retabulating the original returns, he or she must investigate the data collection process at each census used. Alternatively, the researcher can draw upon electronic or published resources which disseminate longitudinal data, e.g., use the data files from IPUMS or HSUS, where the researcher can be assured that the data compilers have addressed the methodological issues involved in preparing the longitudinal file. Even then, the researcher should thoroughly review the technical reports accompanying the data to make sure that any unforeseen issues do not plague the data analysis.



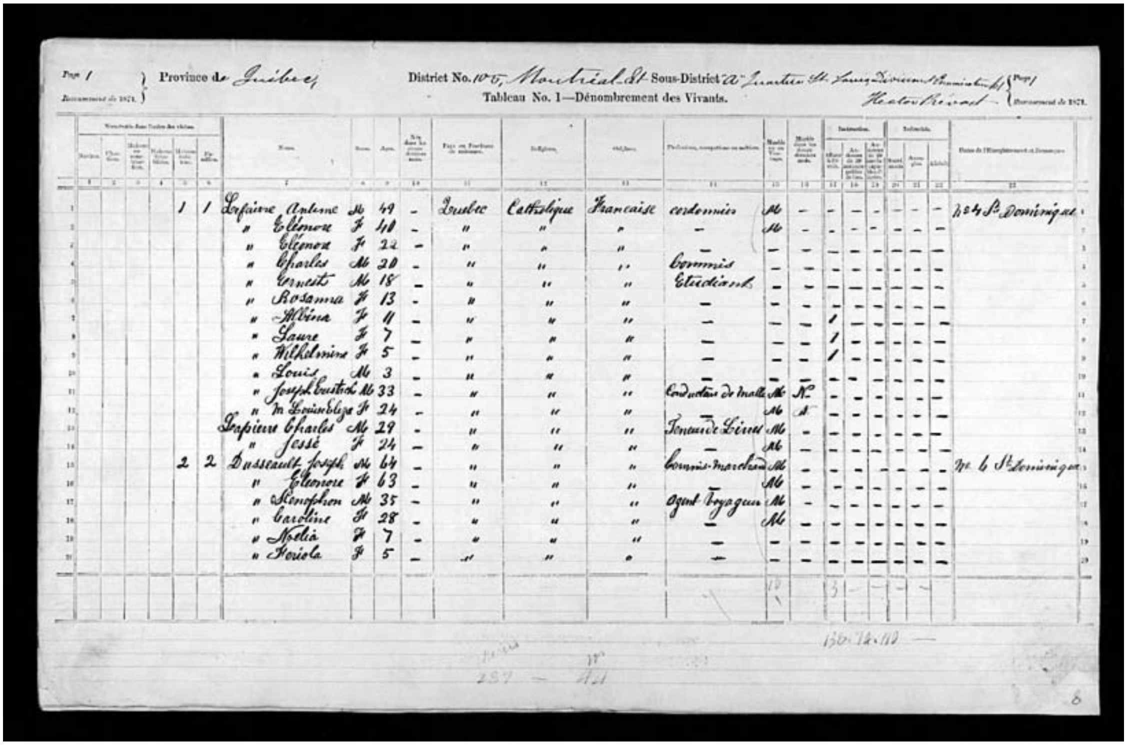
Page 9 Census of 1871. } Province of Ontario District No. 77 Ottawa City. Sub-District C St George's Ward {Page 9 Census of 1871.

Schedule No. 1.—Nominal Return of the Living. Division No 2 [illegible] Enumerator

	1	2	3	4	5	6	7 Names	8 Sex	9 Age	10	11	12 Religion	13 Origin	14	15	16	17	18	19	20	21 Blind	22	23 Dates of Operations and Remarks
1							Ennis Jeane	F	6	—	Q	Catholic	Irish	—	—	—	—	—	—	—	—	—	
2							" Mary	F	4	—	O	"	"	—	—	—	—	—	—	—	—	—	
3							" Alexander	M	2	—	"	"	"	—	—	—	—	—	—	—	—	—	
4							" Edward	M	1/12	Jany	"	"	"	—	—	—	—	—	—	—	—	—	
5							Gaebor Claire Jaché	F	71	—	Q	"	French	—	W	—	—	—	—	—	—	—	
6							Paquette Adeline	F	32	—	"	"	"	Servant	—	—	—	/	/	—	—	—	
7							Ohay Victoria	F	16	—	O	"	Irish	"	—	—	—	—	—	—	—	—	
8					34	36	Quinn John	M	19	—	New Foundland	"	"	Groom	—	—	—	—	—	—	—	—	
9					35	37	Berry Richard	M	21	—	England	Church of England	English	Carpenter	M	—	—	—	—	—	—	—	
10							" Mary	F	28	—	"	"	"	—	"	—	—	—	—	—	—	—	
11					36	38	Green John	M	60	—	"	"	"	Gardener	M	—	—	—	—	—	—	—	
12							" Maria	F	55	—	"	"	"	—	"	—	—	—	—	—	—	—	
13					37	39	Smithson John	M	28	—	Q	Free church Sco	Scotch	Blacksmith	M	—	—	—	—	—	—	—	
14							" Martha	F	24	—	"	"	"	—	"	—	—	—	—	—	—	—	
15							" Agnes	F	1/12	Decr	O	"	"	—	—	—	—	—	—	—	—	—	
16				1	38	40	Schwitzer William	M	26	—	"	Church of England	German	Policeman	M	—	—	—	—	—	—	—	6th April [illegible]
17							" Elizabeth	F	22	—	Q	"	Irish	—	"	—	—	—	—	—	—	—	Enumerator
18							" John Edward	M	1/12	May	O	[illegible]	German	—	—	—	—	—	—	—	—	—	
19							Holbrook Thomas	M	30	—	England	"	English	Bricklayer	M	—	—	—	—	—	—	—	
20							" Eliza	F	20	—	"	"	"	—	"	—	—	—	—	—	—	—	

Figure 2.8 1871 Canadian census form, English

Source: Canadian Geneology Centre, Library and Archives Canada, Census, http://www.collectionscanada.ca/the-public/005-6040-e.html



Page 1 Recensement de 1871. Province de Québec

District No. 105, Montréal-Est Sous-District 2e Quartier St Louis Division [illegible] Heaton Renaud Page 1 Recensement de 1871.

Tableau No. 1—Dénombrement des Vivants.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1					1	1	Lafaivre Antoine	M	49	-	Québec	Catholique	Française	cordonnier	M	-	-	-	-	-	-	-	204 St Dominique
2							" Eléonore	F	40	-	"	"	"	-	M	-	-	-	-	-	-	-	
3							" Eléonore	F	22	-	"	"	"	-	-	-	-	-	-	-	-	-	
4							" Charles	M	20	-	"	"	"	Commis	-	-	-	-	-	-	-	-	
5							" Ernest	M	18	-	"	"	"	Etudiant	-	-	-	-	-	-	-	-	
6							" Rosanna	F	13	-	"	"	"	-	-	-	-	-	-	-	-	-	
7							" Albina	F	11	-	"	"	"	-	-	-	/	-	-	-	-	-	
8							" Laure	F	7	-	"	"	"	-	-	-	/	-	-	-	-	-	
9							" Wilhelmine	F	5	-	"	"	"	-	-	-	/	-	-	-	-	-	
10							" Louis	M	3	-	"	"	"	-	-	-	-	-	-	-	-	-	
11							" Joseph Eustache	M	33	-	"	"	"	Conducteur de Malle	M	N	-	-	-	-	-	-	
12							" M Louise Eliza	F	24	-	"	"	"	-	M	N	-	-	-	-	-	-	
13							Lapierre Charles	M	29	-	"	"	"	Teneur de Livres	M	-	-	-	-	-	-	-	
14							" Jessé	F	24	-	"	"	"	-	M	-	-	-	-	-	-	-	
15					2	2	Dusseault Joseph	M	64	-	"	"	"	Commis-marchand	M	-	-	-	-	-	-	-	No 6 St Dominique
16							" Eléonore	F	63	-	"	"	"	-	M	-	-	-	-	-	-	-	
17							" Xenophon	M	35	-	"	"	"	Agent Voyageur	M	-	-	-	-	-	-	-	
18							" Caroline	F	28	-	"	"	"	-	M	-	-	-	-	-	-	-	
19							" Noelia	F	7	-	"	"	"	-	-	-	-	-	-	-	-	-	
20							" Floriola	F	5	-	"	"	"	-	-	-	-	-	-	-	-	-	
															12		3	-	-	-	-		

130·74·110 —

287 — 411

6

Figure 2.9 Canadian census form, French

Source: Canadian Geneology Centre, Library and Archives Canada, Census, http://www.collectionscanada.ca/geneology/the-public/005-6040-e.html

11 A final note

This essay has leaned heavily on examples from American census practices to illustrate methodological issues. It has done so because there is a particularly rich tradition of unbroken census data collection in the United States, and thus not surprisingly a rich tradition of research using historical census data, both aggregate and micro level, as well as data development projects which have confronted in systematic fashion the technical issues of historical census research.

Similar data traditions and research exist in Scandinavian nations, though the "census" is a different type of data collection. The United Kingdom (and its former colonies) and Canada have also pioneered in historical census use, though their relatively more restrictive rules on access to microdata and interests in demographic and economic research in the pre-census era have led to the development of different types of data sources, particularly parish registers and trade records of economic activity.

Nations without stable governing regimes over long periods, or nations with changing political boundaries, present different challenges for compiling and using longitudinal census data. National regime change tends to disrupt or change the census data collection and reporting process, making the compilation of aggregate data series more difficult, and the preservation of microdata precarious.

Demographic historians, for example, have recently begun to find and use census data collected during the Soviet regime.[3] Researchers are directed to projects which compile international historical statistics generally (e.g., Mitchell, 2003), and to the United Nations Statistical Office.[4]

Broadly speaking, the potential for further development of longitudinal census data, both aggregate and microdata files, is large. As noted above, the computer revolution has made it feasible to retrieve and retrofit data preserved on paper or microfilm to electronic formats. That capacity, in turn, has made it efficient for researchers to develop large-scale historical and comparative census projects, like the North American Population Project, http://www.nappdata.org/napp/, and IPUMS International. Thus, though it may seem counterintuitive, one can expect an increasing flow of data from the pre-electronic past to become available in the years ahead, and thus researchers should be encouraged to think about how their analyses could be enhanced if existing, but inaccessible, historical census data were added to the resources of the social sciences. Envisioning more information from censuses in the past (my apologies to Edward Tufte) is the first step in producing the data.

References

Anderson, M. (1988). *The American Census: A Social History*. New Haven: Yale University Press.

Anderson, M. and Fienberg, S.E. (2001). *Who Counts? The Politics of Census Taking in Contemporary America*, rev. edn. New York: Russell Sage Foundation.

Carter, S. B., et al. (eds) (2006). *Historical Statistics of the United States*, 5 volumes. New York: Cambridge University Press.

Cohen, P. C. (1982). *A Calculating People: The Spread of Numeracy in Early America*. Chicago: University of Chicago Press.

Conk, M. A. (1980).*The United States Census and Labor Force Change: A History of Occupation Statistics*. Ann Arbor, MI: UMI Research Press.

Curtis, B. (2001). *The Politics of Population: State Formation, Statistics, and the Census of Canada, 1840–1875*. Toronto: University of Toronto Press.

Davis, F. J. (1991). *Who Is Black: One Nation's Definition*. University Park, PA: Pennsylvania State University Press.

Ferrie, J. (1996). A new sample of males linked from the public-use microdata sample of the 1850 US Federal Census of Population to the 1860 US Federal Census manuscript schedules. *Historical Methods*, 29: 141–156.

Floud, R. (1979). *An Introduction to Quantitative Methods for Historians*, 2nd edn. New York: Methuen.

Gregory, J. (2005). *The Southern Diaspora: How the Great Migrations of Black and White Southerners Transformed America*. Chapel Hill: University of North Carolina Press.

Guest, A. (1987). Notes from the National Panel Study: Linkage and migration in the late nineteenth century. *Historical Methods*, 20: 63–77.

Hacker, J.D. (2003). Rethinking the "early" decline of marital fertility in the United States. *Demography*, 40: 605–620.

Hall, P. K. and Ruggles, S. (2004). Restless in the midst of their prosperity: New evidence on the internal migration of Americans, 1850–2000. *Journal of American History*, 91: 829–846.

Higgs, E. (1996). *A Clearer Sense of the Census: The Victorian Censuses and Historical Research*. London: HMSO.

Knights, P. R. (1991a). *Yankee Destinies: The Lives of Ordinary Nineteenth-century Bostonians*. Chapel Hill: University of North Carolina Press.

Knights, P. R. (1991b). Potholes on the road of improvement? Estimating census underenumeration by longitudinal tracing: US

[3] See for example, Bakhtior Islamov, "Central Asian Population in Historical Perspectives, "http://www.ier.hit-u.ac.jp/COE/Japanese/Newsletter/No.14.english/Islamov.htm, accessed 11/22/2006; Elena Glavatskaya, "Ethnic categories in the 1926 Russian census," http://www.ddb.umu.se/cbs/workshop2006/abstract.htm #glavatskaya; David Anderson and Konstantin Klokov, "The 1926 Siberian Polar Census and Contemporary Indigenous Land Rights in Western Siberia," http://www.abdn.ac.uk/anthropology/polarcensuslandrights.shtml.

[4] United States Statistics Division, Demographic and Social Statistics, Population and Housing Censuses, http://unstats.un.org/unsd/demographic/sources/census/default.aspx

censuses, 1850–1880. *Social Science History*, 15: 517–526.

Mitchell, B. R. (2003). *International Historical Statistics: Europe, 1750–2000*. London: Palgrave Macmillan.

New York City Department of City Planning (accessed December 5, 2006). *Change in Total Population, 1990 and 2000, New York City and Boroughs*. http://www.ci.nyc.ny.us/html/dcp/html/census/pop2000.shtml.

Nobles, M. (2000). *Shades of Citizenship: Race and the Census in Modern Politics*. Stanford, CA: Stanford University Press.

Nugent, W. (1981). *Structures of American Social History*. Bloomington, IN: Indiana University Press.

Perlmann, J. and Waters, M. C. (eds) (2002). *The New Race Question: How The Census Counts Multiracial Individuals*. New York: Russell Sage Foundation.

Rodriguez, C. (2000). *Changing Race: Latinos, the Census and the History of Ethnicity*. New York: New York University Press.

Rosenoff, T. (2003). *Realignment: The Theory that Changed the Way We Think about American Politics*. New York: Rowman & Littlefield.

Rosenwaike, I. (1972). *Population History of New York City*. Syracuse, NY: Syracuse University Press.

Ruggles, S (2003a). Linking historical censuses: A new approach. IMAG Workshop Montreal, November 10–11, 2003, http://www.pop.umn.edu/research/progress.shtml.

Ruggles, S. (2003b). Multigenerational families in nineteenth-century America. *Continuity and Change*, 18: 139–165.

Szreter, S. (1996). *Fertility, Class and Gender in Britain, 1860–1940*. Cambridge: Cambridge University Press.

Thernstrom, S. (1964). *Poverty and Progress: Social Mobility in a Nineteenth Century City*. Cambridge, MA: Harvard University Press.

US Bureau of the Census (1973). *Population and Housing Inquiries in US Decennial Censuses, 1790–1970*. Washington, DC: GPO.

US Bureau of the Census (1976). *Historical Statistics of the United States, Bicentennial Edition*. Washington, DC: GPO.

US Bureau of the Census (1979). *Twenty Censuses*. Washington, DC: GPO.

Zuberi, Tukufu (2001). *Thicker Than Blood: How Racial Statistics Lie*. Minneapolis: University of Minnesota Press.

This page intentionally left blank

Chapter 3

Repeated cross-sectional research: the general social surveys

Tom W. Smith

1 Introduction

The National Data Program for the Social Sciences (NDPSS) is a social indicators and data diffusion program. Its basic purposes are (1) to gather and disseminate data on contemporary American society in order to (a) monitor and explain change and stability in attitudes, behaviors, and attributes and (b) examine the structure and functioning of society in general as well as the role played by various subgroups; (2) to compare the United States to other societies in order to (a) place American society in comparative perspective and (b) develop cross-national models of human society; and (3) to make high-quality data easily accessible to scholars, students, policymakers, and others with minimal cost and waiting.

These purposes are accomplished by the regular collection and distribution of the National Opinion Research Center (NORC) General Social Survey (GSS) and its allied surveys in the International Social Survey Program (ISSP). Both the GSS and the ISSP surveys have been efficiently collected, widely distributed, and extensively analyzed by social scientists around the world.

2 Organization

The NDPSS is directed by James A. Davis, Tom W. Smith, and Peter V. Marsden. From 1972 to 1982 the GSS was advised by a Board of Advisors and starting in 1978 a Board of Methodological Advisors. In 1983 at the behest of the National Science Foundation (NSF) these groups were combined to form a new Board of Overseers. The Board provides guidance to the GSS, forms linkages to the various research communities, spearheads the development of topical modules, approves the content of each survey, and evaluates the work of the project.

3 Data collection: 1972–2004

Since 1972 the GSS has conducted 25 independent, cross-sectional surveys of the adult household population of the United States and in 1982 and 1987 carried out oversamples of Black Americans. As Table 3.1 details, there have been a total of 45,803 respondents interviewed from the cross-sections, plus 707 Black respondents from the two oversamples.

While the population sampled has remained constant, transitional sample designs have been employed three times: in 1975–76 to calibrate the shift from the original block-quota sample to the full-probability design utilized since 1977, in 1983 when the 1970 NORC sample frame was compared with the new NORC sample frame based on the 1980 census, and in 1993 when the 1980 NORC sample frame and the new 1990 NORC sample frame based on the 1990 census were used. The 1990 sample frame was utilized through 2002. A new sample frame based on the 2000 census was introduced in 2004 (Davis, Smith, and Marsden, 2005).

By using a strict, full-probability sample design, rigorous field efforts, and extensive quality control, the GSS produces a high-quality, representative sample of the adult population of the United States. The GSS response rate has generally been in the upper 70s, with a high in 1993 of 82.4%. However, the GSS response rate has declined in recent years to just over 70%. This rate is higher than that achieved by other major social science surveys and 35–45 percentage points higher than the industry average (Council for Marketing and Opinion Research, 1998; Krosnick, Holbrook, and Pfent, 2003).

In order to accommodate more questions, the GSS employs a questionnaire design under which most questions are asked of only a subset of respondents. From 1972 to 1987, that was accomplished with a rotation design under which questions appeared on two out of every three years. In 1988, the GSS switched from an across-survey rotation design to a split-ballot design. Under this design questions are asked every year, but only on two of three subsamples. Over a three-year period, questions that would have appeared on two surveys with a total of 3000 respondents (2 × 1500) under the old rotation design, now appear on two-thirds subsamples on all three surveys for a total of 3000 respondents (3 × 1000). This shift eliminated the problem of periodic gaps in the annual time series and facilitated time-series analysis (Davis, Smith, and Marsden, 2005).

Starting in 1994, GSS switched to a biennial, double-sample design. In effect the 1994 GSS was two surveys in one with an A sample of 1500 representing the “regular” 1994 GSS and a B sample of 1500 representing the “missing” 1995 GSS. The double-sample design literally combines two separate GSSs with distinct topical and ISSP modules into one field operation (and similarly for the subsequent pairs of years).

3.1 Components

The GSS is divided into five components: (1) the replicating core, (2) topical modules, (3) cross-national modules, (4) experiments, and (5) reinterviews and follow-up studies. In recent years the replicating core has taken up half of the interviewing time and the topical, cross-national, and supplemental modules take up the other half. Experiments are done within either the core or the modules, and reinterviews and follow-up studies involve additional interviewing after the GSS has been completed.

Replicating core

The replicating core consists of questions that regularly appear in surveys *either* as full-coverage items or on subsamples. The content of the core is periodically reviewed by the PIs and Board of Overseers to insure that the content remains relevant and up-to-date. Currently, the replicating core makes up about half of the overall length of the GSS and consists of about one-third demographic questions and two-thirds attitudes and behaviors. The replicating core forms the basis for the trend analysis and pooling of cases for subgroup analysis.

The GSS is intentionally wide-ranging in its contents, with 4624 variables in the 1972–2004 cumulative file. One needs to peruse the GSS Cumulative Codebook (Davis, Smith, and Marsden, 2005) or the online version at http://www.icpsr.umich.edu/cgi-bin/bob/newark?study=4295 to fully appreciate the scope of the GSS.

The GSS is different from most surveys in the wide variety of demographics included and the detail in which they are asked and coded. In addition to covering the extensive background variables on the respondent's current status, the GSS has extensive information on the respondent's family of origin and parental characteristics. Among the family of origin items are questions on the intactness of families (and reasons for "broken homes"), number of siblings, religion, region, and community type. Additionally, parental variables include mother's and father's education, church attendance, occupation, and industry. There are also many questions about spouses.

In addition, measures are usually very detailed. For example, occupation and industry use both the census three-digit classification codes and the four-digit International Standard Classification of Occupations, two measures of occupational prestige, education codes both number of years in school and highest degree obtained, three community type measures are included, and up to three ethnic and racial identities are coded.

Besides the demographics, the core items cover a variety of behaviors, personal evaluations, and attitudes about central social and political issues from death (e.g., capital punishment, suicide, euthanasia) to taxes (as a redistribution measure, paying too much?). Among the many topics covered are abortion, civil liberties, confidence in institutions, crime and punishment, government-spending priorities, poverty and inequality, intergroup relations, religion, and women's rights.

Topical modules

Topical modules (special sections on a particular theme) first appeared in 1977 and have been an annual feature since 1984. The topical modules are designed to facilitate both innovation and greater depth. They introduce new topics not previously investigated by the GSS and cover existing topics in greater detail with more fully-specified models. The original concept for a module may come from the principal investigators, the Board of Overseers, or other interested scholars. The themes covered in major modules are listed in Table 3.1.

Cross-national modules

The GSS has spurred cross-national research by inspiring other nations to develop similar data-collection programs (e.g., the ALLBUS (Germany), British Social Attitudes, National Social Science Survey (Australia), Taiwan Social Change Study, Polish General Social Survey, Japanese General Social Survey, Korean General Social Survey, and Chinese General Social Survey (Smith, Koch, Park, and Kim, 2006b) and by organizing these and other programs into the ISSP. (See www.issp.org)

The fundamental goal of ISSP is to study important social and political processes in comparative perspective. In addition, by replicating earlier modules, ISSP not only has a cross-national perspective, but also an over-time perspective. With ISSP one can both compare nations and test whether similar social-science models operate across societies, and also see if there are similar international trends and whether parallel models of societal change operate across nations. Thus, by combining an across-time with a cross-national design, ISSP incorporates two powerful perspectives for studying societies.

ISSP evolved from a bilateral collaboration between the Allgemeinen Bevolkerungsumfragen der Sozialwissenschaften (ALLBUS) of the Zentrum fuer Umfragen, Methoden, und Analysen (ZUMA) in Mannheim, West Germany and the GSS of NORC, University of Chicago. In 1982 and 1984 ZUMA and NORC devoted a small segment of the ALLBUS and GSS to a common set of questions on job values, important areas of life, abortion, feminism, class differences, equality, and the welfare state.

Meanwhile, in late 1983 the National Centre for Social Research, then known as Social and

Table 3.1 Design features of the GSS 1972–2004

Year	*Sample size*	*Sample type*	*Response rate%*	*Item rotation*	*Experimental forms*	*Reinterviews*	*Modules topical*	*International*
1972	1613	BQ	–	None	None	Two waves	None	None
1973	1504	BQ	–	AS	Two forms	Three waves	None	None
1974	1484	BQ	–	AS	Two forms	Three waves	None	None
1975	1490	1/2BQ 1/2FP	– 75.6	AS	Split sample	None	None	None
1976	1499	1/2BQ 1/2FP	– 75.1	AS	Two forms + split sample	None	None	None
1977	1530	FP	76.5	AS	None	None	Race, abortion, feminism	None
1978	1532	FP	73.5	AS	Two waves	None	None	None
1980	1468	FP	75.9	AS	Three forms	None	None	None
1982	1506	FP	77.5	AS	Two forms	None	Military	ZUMA
1982B	354	FP	71.7	AS	Two forms	None	Military	ZUMA
1983	1599	70FP 80FP	79.4	AS	Two forms + split sample	None	None	ZUMA
1984	1473	FP	78.6	AS	Three forms	None	None	ZUMA
1985	1534	FP	78.7	AS	Two forms	None	Social networks	ISSP

1986	1470	FP	75.6	AS	Two forms + vignettes	None	Welfare	ISSP
1987	1466	FP	75.4	AS	Three forms	Political tolerance	Political participation	ISSP
1987B	353	FP	79.9	AS	Three forms	Political tolerance	Political participation	ISSP
1988	1481	FP	77.3	SB	Two forms	Cognitive	Religion	ISSP
1989	1537	FP	77.6	SB	Two forms[a]	Methods/ Health[b]	Occupational prestige	ISSP
1990	1372	FP	73.6	SB	Two forms	Health	Intergroup relations	ISSP
1991	1517	FP	77.8	SB	Two forms 1992	ISSP	Work organizations	ISSP
1993	1606	FP	82.4	SB	Two forms	None	Culture	ISSP
1994	2992	FP	77.8	DSB	Two forms	None	Family mobility Multiculturalism	ISSP
1996	2904	FP	76.1	DSB	Two forms + vignettes	Parents of Students	Mental health Emotions Gender Market exchange	ISSP
1998	2832	FP	75.6	DSB vignettes	Two forms + knowledge	Health use	Religion Job experiences Health and mental health Medical ethics Culture Inter-racial friendships	ISSP

(*Continued*)

Table 3.1 (Continued)

Year	*Sample size*	*Sample type*	*Response rate%*	*Item rotation*	*Experimental forms*	*Reinterviews*	*Modules topical*	*International*
2000	2817	FP	70.0	DSB	Two forms	Internet use	Religion Computers Multi-ethnic Health status Freedom	ISSP
2002	2765	FP	70.1	DSB	Two forms	Worker health	Altruism Internet Intergroup relations Quality of work Worker pay Adulthood Doctors Mental health The arts	ISSP
2004	2817	FP	70.4	DSB	Two forms	Voluntary associations	Altruism Internet Negative events Genes/Environment Religious change Guns Social networks/ Voluntary groups Alcohol use Workplace stress/ Violence	ISSP

[a] For the OCCUPATIONAL PRESTIGE module 12 subsamples were used.
[b] The 1990 health reinterview used 1989 and 1990 GSS respondents.

B = Black oversample
BQ = Block quota sampling
FP = Full probability sampling
AS = Across-survey rotation
SB = Split-ballot rotation
DSB = Double sample, split-ballot rotation

Community Planning Research (SCPR), which was starting the British Social Attitudes Survey, secured funds for meetings for international collaboration. Representatives from ZUMA, NORC, SCPR, and the Research School of Social Sciences, Australian National University, organized ISSP in 1984 and agreed to (1) jointly develop topical modules covering important social science topics, (2) field the modules as supplements to the regular national surveys (or a special survey if necessary), (3) include an extensive common core of background variables, and (4) make the data available to the social-science community as soon as possible.

Each research organization funds all of its own costs. There are no central funds. Coordination is supplied by one nation serving as the secretariat. The United States served as the secretariat from 1997 to 2003.

Since 1984, ISSP has grown to 40 nations, the founding four—Germany, the United States, Great Britain, and Australia—plus Austria, Brazil, Bulgaria, Canada, Chile, Croatia, Cyprus, the Czech Republic, Denmark, the Dominican Republic, Finland, Flanders, France, Hungary, Ireland, Israel, Japan, Korea, Latvia, Mexico, the Netherlands, New Zealand, Norway, the Philippines, Poland, Portugal, Russia, Slovakia, Slovenia, South Africa, Spain, Sweden, Switzerland, Taiwan, Uruguay, and Venezuela. In addition, East Germany was added to the German sample upon reunification. Past members not currently active include Bangladesh and Italy. In addition, a number of non-members have replicated one or more ISSP modules. This includes Argentina (Buenos Aires metro area only), Lithuania, and Singapore.

ISSP maintains high standards of survey research. Each nation uses full-probability sampling, carefully monitors all phases of the data collection, and cleans and validates the data. The ISSP's Central Archive further checks all data archived by the member nations. Countries applying for membership answer a series of standard questions about methodology and survey procedures. Only once the secretariat has received satisfactory responses to all questions is a country's membership application considered by ISSP. Each country reports to ISSP its methods and various technical details such as its response rate. To check on the representativeness of the sample, each country compares distributions on key demographics from ISSP surveys to the best data sources in their respective countries.

ISSP modules have covered the following topics: (1) Role of Government—1985, 1990, 1996, 2006,[1] (2)Social Support and Networks (1986 and 2001), (3) Social Inequality (1987, 1992, 1999), (4) Gender, Family, and Work (1988, 1994, 2002), (5) Work Orientation (1989, 1997, 2005), (6) Religion (1991, 1998, 2008), (7) Environment (1993, 2000), (8) National Identity (1995, 2003), (9) Citizenship (2004), and (10) Leisure Time (2007).

Experiments

Experimental forms have always been a regular part of the GSS. The GSS has used split samples in 1973, 1974, 1976, 1978, 1980 and 1982–2004. They have been an integral part of the GSS's program of methodological research. Dozens of experiments have examined differences in question wording, response categories, and context (Davis, Smith, and Marsden, 2005).

Experiments are carried out as part of the replicating core, topical modules, and supplements. In some years the experiments consist of additional questions not regularly appearing on the GSS, such as the interracial friendships experiments in 1998 and the wording and response-order experiments on genetic screening items in 1991 and 1996. Most of the time, however, the experiments compare a variant wording or order with the standard GSS wording and/or order being the control. Examples

[1] ISSP replication modules repeat two-thirds of their content from earlier rounds.

are the experiments on measuring race and ethnicity in 1996 and 2000.

In addition, there have often been experiments within topical modules. For example, experiments were conducted as part of the 1986 factorial-vignette study of welfare, the occupational-prestige study in 1989, the 1989 intergroup-relations module with wording experiments to test the impact of class versus racial references, the 1994 multiculturalism module with various formulations of affirmative action policies, the 1996 mental-health module with 18 different versions of five basic vignettes (90 versions in all) to examine stigmatization of troubled individuals, the 1996 gender module, the 1998 factorial vignettes on terminal-care decisions, the 2000 health-status and computer-use modules, and the 2002 vignette studies of the mental health of children and physician–patient communications.

Reinterviews and follow-ups

GSS respondents have been reinterviewed both as part of methodological and substantive studies. The methodological uses have included studies of reliability, cognition, and wording and context. In 1972, 1973, 1974, and 1978, test/retest studies of item stability and reliability were conducted (Smith and Stephenson, 1979; Alwin and Krosnick, 1989). In 1988, cognitive scientists at the University of Chicago expanded the normal GSS validation effort and added recall questions about the timing and content of the initial interview. Reinterview reports were then validated against the known information on date and content and models of memory were developed to explain the discrepancies. Telescoping or forward biasing in the reporting of past events was documented and this was related to the placing of upper limits on time estimates and a tendency to round to the next lower or complete time period, e.g., two weeks, one month (Huttenlocher, Hedges, and Bradburn, 1990).

In 1990, NORC and the University of Chicago supported a seminar on survey-research methods to study wording and context effects. About a third of the 1989 GSS cases were recontacted by phone. Comparisons were made between standard and variant questions across subsamples on the reinterview, between standard questions on the GSS and the reinterview, and between standard questions on the GSS and variant questions on the reinterviews. As in the earlier GSS reinterview studies, a notable degree of instability in responses was found (Junn and Nie, 1990; Ramirez, 1990). As expected, attitudinal items showed more variation than demographics. The less educated, those with no earned income, and older respondents showed the greatest differences in their responses.

The GSS has also served as a list sample for several substantive studies. GSS respondents are a representative sample of adults living in households and can be used as a list or sample frame for a follow-up study. While one must naturally adjust for any bias from panel mortality, the GSS offers an excellent frame for a follow-up study. First of all, since respondent names, addresses, and telephone numbers are known, GSS respondents are relatively easy to recontact. Second, a rich amount of information is known about respondents. This information can be used in several ways. For unchanging attributes like year of birth, income during the past year, or nationality, one can link the data obtained on the GSS to the follow-up study and thereby free up time on the follow-up study. Third, one can use any GSS variables to study panel mortality and, if necessary, adjust for panel mortality bias.

There have been seven substantive reinterviews of GSS respondents. The first in 1987 contained questions on political tolerance and Cloninger's Tridimensional Personality Scale. The second reinterview study was the 1990 National Survey of Functional Health Status.

Respondents from the 1989/1990 GSS, plus an additional sample of people 65+ from these households, were contacted in late 1990 and early 1991. In 1994–95, respondents were reinterviewed again. In the third reinterview study in 1992 respondents to the 1991 GSS were reinterviewed in order to collect information for the ISSP social inequality module and study changes in negative life events over time. The fourth reinterview in 1997 contacted parents of students in grades 1–8 from the 1996 GSS. The fifth on the 1998 GSS did reinterviews on knowledge about and attitudes towards the role of behavioral interventions and social-science treatments in health care. The sixth reinterview in 2001 was an extension of the 2000 topical module on computers and the Internet. The latest reinterview is of employed people on the 2002 GSS. In 2002–2003 they were reinterviewed about work-related, health issues.

The GSS has also served as the source for six special follow-up studies, most involving hypernetwork sampling. First, in 1991 a record of the employer of respondents and spouses was collected. These employers were contacted as part of a study of work organizations, the National Organizations Study (NOS). This information can be analyzed in its own right as well as linked back to the attitudes of the original GSS respondents. Second, in 1994 a random sibling was selected for an interview in order to study social mobility within sibsets. Third, in 1998 and 2000 a sample of respondents' congregations was created. In 1998 a follow-up survey of these congregations was fielded. For 2000 there were follow-up surveys both of congregations and of people attending services of these congregations. Fifth, as with the 1991 NOS, on the 2002 GSS information was collected on respondents' employers (spouses' employers were not covered in 2002). Finally, in 2006 the National Voluntary Associations Study contacted groups that 2004 GSS respondents belonged to.

4 Publications by the user community

As of 2005, the GSS was aware of over 12,000 research uses of the GSS in articles, books, dissertations, etc. Most users (82%) have been academics with college affiliations. Other users include scholars at research centers, foundations, and related organizations (12%); government researchers (1%); and others and unknown (5%). Among the academics sociologists predominate (56%), followed by political scientists (15%), law and criminal justice researchers (6%), psychologists (5%), economists (4%), physicians and other health professionals (5%), statisticians (3%), business management professors (2%), other social scientists (e.g. anthropologists and geographers) (2%), and non-social scientists and miscellaneous (2%).

Moreover, with the exception of the census and its Current Population Survey, the GSS is the most frequently used dataset in the top sociology journals.[2] As Table 3.2 shows, in the top sociology journals the GSS has been used in 145

Table 3.2 Most frequently used datasets in leading sociology journals, 1991–2003

Census/CPS	180
GSS	145
National Longitudinal Survey of Youth	43
Panel Survey of Income Dynamics	36
National Survey of Families and Households	28
National Educational Longitudinal Survey	18
Adolescent Health	14
High School and Beyond	13
National Election Studies	13
Occupational Change in a Generation II	10

[2] we used the *American Sociological Review*, the *American Journal of Sociology*, and *Social Forces*. They are the consensus choice as the top general sociological journals (Allen, 1990; Kamo, 1996; Presser, 1984).

articles—more often than the total of the next five most frequently used datasets combined.

5 Teaching and other uses

The GSS is widely used in teaching at the undergraduate and graduate levels. About 250,000 students annually take courses that utilize the GSS and nearly 400 college textbooks use GSS data.

The GSS has also been used outside the academic community by the government, media, non-profits, and business community. Taking the federal government as an example, the GSS is regularly used by (1) the Congressional Reference Service of the Library of Congress, (2) the *Science and Engineering Indicators* series of NSF, (3) the *Sourcebook of Criminal Justice Statistics* of the Bureau of Justice Statistics, and (4) *Statistical Abstract of the United States* of the Bureau of the Census. GSS data have been cited in 20 briefs to the US Supreme Court.

6 Contributions to knowledge

Because of the wide-ranging content and extensive level of usage of the GSS, it is effectively impossible to describe all of the results from the thousands of publications covering dozens of fields. Instead GSS's contributions to basic knowledge will be considered regarding (1) how key design features of the GSS have promoted social-science research, (2) the study of societal change, (3) cross-national research, and (4) methodological research.

6.1 Design features of the GSS and research

Several key aspects of the GSS study design greatly facilitate research opportunities. These include: (1) replication, (2) breadth of substantive content, (3) extensive and detailed demographics, (4) providing a standard of comparison for other surveys, and (5) depth and innovation in the topical modules.

Replication is the most important design feature of the GSS. Replication is necessary for two crucial research goals of the GSS: (1) the study of societal change and (2) the study of subgroups. A sample of GSS research publications since 1995 shows that 60% of all GSS usages make use of the replication feature by utilizing two or more years of the GSS.

The GSS core is based on the simple principles that (1) the way to measure change is not to change the measure (Smith, 2005b), and (2) the optimal design for aggregating cases is a replicating cross-section. Besides replication within the core to study societal change and subgroups, the GSS employs replication in several other ways.

First, many of the variables used on the GSS were adopted from baseline surveys with observations going back as far as the 1930s and 1940s. As a result, hundreds of GSS trends extend back before the inception of the GSS in 1972 (Smith, 1990).

Second, several topical modules have been designed to replicate seminal studies. For example, the 1987 module on sociopolitical participation replicated key segments of the 1967 Verba–Nie study of political participation (Verba and Nie, 1972); the 1989 occupational prestige module updated the NORC prestige studies of 1963–1965 (Nakao and Treas, 1994); and the 1996 Mental Health module drew on Starr's seminal study from the early 1950s (Phelan, et al., 2000). Even when not primarily a replication, other modules, such as the modules in 1990 on intergroup relations, in 1991 on work organizations, in 1994 on multiculturalism, in 2000 on health functioning, and in 2002 and 2004 on empathy and altruism, have adopted key scales from earlier studies.

Third, there is a social trends component in ISSP. Cross-national modules are periodically repeated to measure societal change in a comparative perspective.

Finally, experiments have been replicated over time.

Replication is first and foremost used to study societal change. An analysis of recent publications (from 1995 on) shows that 39% of all research examines trends. Examples of this body of research are presented in the section on research findings below.

Replication is also essential for the pooling of cases to study cultural subgroups and understand the great complexity and diversity of American society. For example, the 1972–2004 GSSs have 973 Jews, 2850 holders of graduate-level degrees (and 850 with 20+ years of schooling), 498 French Canadians, 732 registered nurses, and 72 economists. The GSS has been used not only to study all of the major social groups (e.g., men and women, Blacks and Whites, the employed, etc.), but also to examine much smaller groups and combinations of groups. The GSS has been used to focus on and examine an incredibly wide range of groups, including: American Indians (Hoffman, 1995), art museum attendees (DiMaggio, 1996), engineers and scientists (Smith, 2000; Weaver and Trankina, 1996), farmers (Drury and Tweeten, 1997), part-time workers (Kalleberg, 1995), schoolteachers (Lindsey, 1997), the self-employed and business owners (Kingston and Fries, 1994), and veterans (Feigelman, 1994; Lawrence and Kane, 1995). And among the combination of groups investigated are Black Catholics (Feigelman, Gorman, and Varacalli, 1991), the divorced elderly (Hammond, 1991), older rural residents (Peterson and Maiden, 1993), and self-employed women (Greene, 1993; McCrary, 1994). Moreover, in a number of instances subgroups were pooled into several time periods so that both trend and subgroup analysis was possible. For example, among Hispanics (Hunt, 1999), Jews (Greeley and Hout, 1999; Smith, 2005a), and schoolteachers (Walker, 1997).

A second key design feature of the GSS is its wide-ranging content. The cumulative 1972–2004 GSS dataset has 4624 variables and typically 850–1000 variables appear on each recent GSS. As a result, the GSS covers a wide range of topics and as the Office of Inspector General of NSF has noted, attracts use from "scientists in almost every subfield of sociology and in numerous other social science disciplines (Office of Inspector General, 1994)."

This allows investigators to test hypotheses across a large number of variables rather than being restricted to a handful of items. For example, Davis (2000) looked at trends on 81 items, Freese, Powell, and Steelman (1999) examined birth order differences with 106 variables, Smith (2005a) considered ethnic and religious differences across 150 variables, and Greeley (1995) utilized 230 variables to study religion.

A third key design feature is the GSS's rich and detailed set of demographics. As discussed above, the GSS has backround variables on respondents, spouses, household, and parents, and many multiple measures on such variables as race/ethnicity, occupation, income, and community type.

Finally, the GSS serves as a standard for many other surveys. It is widely used as a national norm for comparison with student, local, state, international, and special samples.

6.2 Societal change

The GSS is the single best source of trends in social attitudes available. The 1972–2004 GSSs have time trends of over 1400 variables with hundreds spanning 30+ years. As Nie, Junn, and Stehlik (1996) have noted, the GSS "is the only continuous monitoring of a comprehensive set of non-economic attitudes, orientations, and behaviors in the United States today." Or as Morin (1998) characterized it, the GSS is "the nation's single most important barometer of social trends."

Many general studies of societal change have been carried out. DiMaggio, Evans, and Bryson found little support for the simple, attitude polarization hypothesis. Most scales and items did not become more polarized under several definitions, but some important, but isolated, examples did emerge (DiMaggio, Evans, and Byson, 1996; DiMaggio and Bryson, forthcoming). Likewise, Hochschild (1995) found convergence regarding the "American Dream" across race and class lines. Smith (1994; 1997) and Davis (1995; 2000) found that most societal change in attitudes is (1) slow, steady, and cumulative, and (2) that most societal change is explained (in decreasing order of importance) by (a) cohort-education turnover models, (b) episodic shocks (e.g., wars and political scandals), and (c) structural changes in background variables.

Many studies of change within particular topics have also been conducted. One of the top areas is social capital. Putnam and others (2000; Crawford and Levitt, 1998) have argued that social capital is eroding and this is seriously undermining the smooth operation of the political system and society in general. Ladd (1996; 1999) counters that the change is both exaggerated and is not so much a decline, but a reconfiguration of civil society. Similarly, Paxton (1999) finds a mixed pattern of change with a decline in individual trust, no general decline in trust in institutions, and no decline in voluntary associations.

Intergroup relations is another major area of analyzing trends. Research indicates that intergroup relations are multidimensional and multiple indicators are needed to track attitudes towards many different aspects (e.g., target groups, principles, policies, role of government, etc.). Schuman and colleagues (Schuman and Krysan, 1999; Schuman, Steeh, Bobo, and Krysan, 1997) have demonstrated that trends have proceeded at very different rates, with quick and large-scale shifts towards the principle of racial equality at one extreme to little or no gain in support for concrete measures to ensure equal treatment at the other end.

Societal changes in family values have also been frequently examined and show a massive shift from traditional to modern attitudes and practices. Smith (1999) showed that many family values have become less traditional and that the changes in family values were both assisted by changes in family structure and in turn facilitated the shift in the composition of households. Popenoe and Whitehead (1999) focused on the declining centrality of marriage over the last generation. Alwin (1996) showed how the coresidence preferences of families changed both across time and across cohorts. Straus and Mathur (1996) found that support for both spanking and obedience in children declined. Brewster and Padavic (1998), Misra and Panigrahi (1995), and Rindfuss, Brewster, and Kavee (1996) isolated gender interaction and cohort effects as the top causes of shifts in gender role attitudes.

Of course the GSS also covers trends in scores of other areas. For example, Davis and Robinson (1998) showed a notable shift in the class identities of married couples with both husbands and wives increasingly using the wives' characteristics in assessing their own class identity. Hunt (1999) indicated that the Hispanics have become less Catholic both across time and across immigrant generations. Since *Occupational Change in a Generation II* in 1973, the GSS has been the main source of data on changes in intergenerational mobility. As Mare (1992) noted, "Except for the NORC General Social Survey (GSS), we have no standard vehicle for monitoring the process of social stratification..." Recent examinations of the trends in mobility include Davis (1994), Hauser (1998) and Hout (1997).

6.3 Cross-national

With 19 completed and released modules and 2335 usages, ISSP has produced a body of research that has been almost as wide-ranging and difficult to summarize as the GSS in general. (For the latest ISSP bibliography see www.issp.org)

As a single example of the cross-national uses, consider the 1995–96 and 2003–2004 national identity modules. They have been used to examine the shifting role of the nation state as its position has been changed both from above by regional and international organizations (e.g., EU, NAFTA, UN, WTO) and from below by movements for autonomy and local self-government, and to determine the cultural identity and distinctiveness of individual countries (e.g., Hjerm, 1998; 2004; Jones, 2001; McCrone and Surridge, 1998; Peters, 2002). For example, Smith and Jarkko (1998) and Smith and Kim (2006) showed that national pride in ten domains was determined by a combination of objective conditions and a people's understanding of their history. They also showed that national pride was uniformly lower among ethnic, racial, religious, linguistic, and regional minorities and that national pride has declined across birth cohorts in almost all countries.

6.4 Methodological research

The GSS gives the highest priority to maintaining data quality and minimizing measurement error. In part this has been carried out by the adoption of rigorous design and execution standards (e.g., full-probability sampling, pretesting and careful item development, maintaining a high response rate, data validation, data cleaning, etc.). In addition, this has been achieved by carrying out one of the most extensive programs of methodological research in survey research. The project has 105 GSS Methodological Reports that use both experimental and non-experimental designs to study virtually all aspects of total survey error (Davis, Smith, and Marsden, 2005). Among the topics covered are: (1) the reliability and validity of behavioral reports; (2) test/retest reliability; (3) sample-frame comparability; (4) sensitive topics; (5) third-person effects; (6) education/age-cohort interactions; (7) nonresponse bias, (8) the measurement of race and ethnicity; (9) context effects; (10) question wording; (11) scale construction; (12) item nonresponse, and (12) cross-national comparisons.

7 Summary

The GSS has aptly been described as a "national resource" (Firebaugh, 1997; Working Group on Large-Scale Data Needs in Luce, Smelser, and Gerstein, 1989), as a "core database" in both sociology and political science (Campbell, 2001; Kasse, 2001), as a "public utility for the community at large" (Office of Inspector General, 1994), as having "revolutionized the study of social change" (ICPSR, 1997), and as "a major source of data on social and political issues and their changes over time" (AAPOR Innovators Award, 2000).

In order to serve the social-science community, the GSS draws heavily upon that community of scholars in the selection and development of modules and items. Between the Board and developmental committees *hundreds* of researchers have participated in the design of GSS components. Then the GSS provides quick, equal, and easy access to the data which in turn leads to widespread utilization of the data by *thousands* of social scientists and *hundreds of thousands* of their students. It is not only widely used in the United States, but especially through ISSP it is used by scholars around the world. The known GSS research usages number over 12,000. Usage has been especially strong in the top sociology journals where only data collected by the Bureau of the Census are used more frequently than the GSS.

In sum, the GSS produces top-quality, representative data for the United States and, through ISSP, in many other countries on topics of fundamental importance to the social sciences, is extremely widely used in both teaching and research, and has considerably expanded the knowledge base in the social sciences in a very cost-effective manner.

References

Allen, M. P. (1990). The “quality” of journal in sociology reconsidered: Objective measurers of journal influence. *ASA Footnotes*, (Nov.), 4–5.

Alwin, Duane F. (1996). Coresidence beliefs in American society: 1973 to 1991. *Journal of Marriage and the Family*, 58: 393–403.

Alwin, Duane F. and Krosnick, Jon A. (1989). The reliability of attitudinal survey data: The impact of questions and respondent characteristics. *GSS Methodological Report No. 61.* Chicago: NORC.

Brewster, Karin L. and Padavic, Irene (1998). Change in gender ideology, 1977–1996: The contributions of intracohort change and population turnover. Paper presented to the American Sociological Association, San Francisco (August).

Campbell, Richard T. (2001). Databases, core: Sociology. In Neil J. Smelser and Paul B. Bates (eds), *International Encyclopedia of the Social and Behavioral Sciences.* New York: Elsevier.

Council for Marketing and Opinion Research (1998). Respondent cooperation. *Opinion*, 1: 2.

Crawford, Susan and Levitt, Peggy (1998). American societal change and civic engagement: The case of the PTO. Harvard University Report.

Davis, James A. (1995). Patterns of attitude change in the USA: 1972–1994. Paper presented to the Conference on “A Decade of Change in Social Attitudes,” London.

Davis, James A. (2000) Testing the demographic explanation of attitude trends: Secular trends in attitudes among US householders, 1972–1996. *Social Science Research*, 30: 363–385.

Davis, James A., Smith, Tom W. and Marsden, Peter V. (2005). *General Social Surveys, 1972–2004: Cumulative Codebook*. Chicago: NORC.

Davis, Nancy J. and Robinson, Robert V. (1998). Do wives matter? Class identities of wives and husbands in the United States, 1974–1994. *Social Forces*, 76: 1063–1086.

DiMaggio, Paul (1996). Are art-museum visitors different from other people? *Poetics*, 24: 161–180.

DiMaggio, Paul and Bryson, Bethany (forthcoming). Americans’ attitudes towards cultural authority and cultural diversity: Culture wars, social closure, or multiple dimensions. In *Contemporary US Democracy*. Washington, DC: Smithsonian Institution Press.

DiMaggio, Paul, Evans, John, and Bryson, Bethany (1996). Have Americans’ social attitudes become more polarized? *American Journal of Sociology*, 102: 690–755.

Drury, Renee and Tweeten, Luther (1991) Have farmers lost their uniqueness? *Review of Agricultural Economics*, 19: 58–90.

Feigelman, William (1994) Cigarette smoking among former military service personnel: A neglected social issue. *Preventive Medicine*, 23: 235–241.

Feigelman, William, Gorman, Bernard S. and Varacalli, Joseph (1991). The social characteristics of Black catholics. *Sociology and Social Research*, 75: 133–143.

Firebaugh, Glenn (1997). *Analyzing Repeated Surveys.* Thousand Oaks, CA: Sage.

Freese, Jeremy, Powell, Brian and Steelman, Lala Carr (1999). Rebel without a cause or effects: Birth order and social attitudes. *American Sociological Review*, 64: 207–231.

Greeley, Andrew M. (1995). *Religion as Poetry: An Empirical Model*. New Brunswick: Transaction.

Greeley, Andrew M. and Hout, Michael (1999). Americans’ increasing belief in life after death: Religious competition and acculturation. *American Sociological Review*, (Dec.): 813–835.

Greene, Patricia G. (1991). A theoretical and empirical study of self-employed women. Unpublished Ph.D. dissertation, University of Texas.

Hammond, Ronald (1998). A model of life satisfaction for the elderly divorced and separated. Unpublished Ph.D. dissertation, Brigham Young University.

Hauser, Robert M. (1998). Intergenerational economic mobility in the United States: Measures, differentials, and trends. Paper presented to the American Sociological Association, San Francisco (August).

Hjerm, Mikael (1998). National identities, national pride, and xenophobia: A comparison of four western countries. *Acta Sociologica*, 41: 335–347.

Hjerm, Mikael (2004). Defending liberal nationalism: At what cost? *Journal of Ethnic and Migration Studies*, 30: 1–17.

Hochschild, Jennifer (1995). *Facing Up to the American Dream: Race, Class, and the Soul of the Nation.* Princeton: Princeton University Press.

Hoffman, Thomas J. (1995). American Indians and the general social survey: 1973–1993. Paper presented to the Western Social Science Association, Oakland (April).

Hout, Michael (1997). Opportunity, change, and restructuring: Women’s and men’s occupational mobility. Paper presented to the American Sociological Association, Toronto (August).

Hunt, Larry L. (1999). Hispanic protestantism in the United States: Trends by decade and generation. *Social Forces*, 77: 1601–1623.

Huttenlocher, Janellen, Hedges, Larry V. and Bradburn, Norman (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology*, 16: 196–213.

ICPSR (1997). ICPSR present awards for outstanding contributions. *ICPSR Bulletin*, 18 (Dec.).

Jones, Frank L. and Smith, Philip (2001). Individual and societal bases of national identity. *European Sociological Review*, 17: 103–118.

Junn, Jane and Nie, Norman (1990). The GSS phone reinterview assessment of response effect. Paper presented to the NORC Seminar series, Chicago.

Kalleberg, Arne L. (1995). Part-time work and workers in the United States: Correlates and policy issues. *Washington and Lee Law Review*, 52: 771–798.

Kamo, Yoshinori (1996). Ranking sociology department: A different perspective. *ASA Footnotes*, 24: 4.

Kasse, Max (2001). Databases, core: Political science. In Neil J. Smelser and Paul B. Bates (eds), *International Encyclopedia of the Social and Behavioral Sciences*. New York: Elsevier.

Kingston, Paul W. and Fries, John C. (1994). Having a stake in the system: The sociopolitical ramifications of business and home ownership. *Social Science Quarterly*, 75: 679–686.

Krosnick, Jon, Holbrook, Allyson and Pfent, Alison (2003). Response rates in surveys by the news media and government contractor survey research firms. Paper presented to the American Association for Public Opinion Research, Nashville.

Ladd, Everett C. (1996). The data just don't show erosion of America's "social capital". *The Public Perspective*, 7: 1, 5–21.

Ladd, Everett C. (1999). *The Ladd Report*. New York: The Free Press.

Lawrence, George H. and Kane, Thomas D. (1995/6). Military service and racial attitudes of White veterans. *Armed Forces and Society*, 22: 235–247.

Lindsey, Stephen D. (1997). Racial and ethnic attitudes of the American teacher and implications for the implementation of a culturally pluralistic philosophy in schools. Unpublished EdD dissertation, Texas A&M University.

Luce, R. Duncan, Smelser, Neil J. and Gerstein, Dean R. (eds) (1989). *Leading Edges in Social and Behavioral Science*. New York: Russell Sage Foundation.

Mare, Robert D. (1992). Trends in the process of social stratification. *Contemporary Sociology*, 21: 654–658.

McCrary, Michael (1994). Gendered labor markets and family resources: Sex differences in access to self-employment. Paper presented to the American Sociological Association, Los Angeles.

McCrone, David and Surridge, Paula (1998). National identity and national pride. In Roger Jowell, et al. (eds), *British and European Social Attitudes*. Aldershot: Ashgate.

Misra, Ranjita and Panigrahi, Bhogahan (1995). Change in attitudes toward working women: A cohort analysis. *International Journal of Sociology and Social Policy*, 15: 1–20.

Morin, Richard (1998). New facts and hot stats from the social sciences. *Washington Post*, June 14, C5.

Nakao, Keiko and Treas, Judith (1994). Updating occupational prestige and socioeconomic scores: How the new measures measure up. *Sociological Methodology*, 24: 1–72.

Nie, Norman H., Junn, Jane and Stehlik-Barry, Kenneth (1998). *Education and Democratic Citizenship in America*. Chicago: University of Chicago Press.

Office of the Inspector General (1994). National Opinion Research Center, Chicago, Illinois: Inspection Report. National Science Foundation.

Paxton, Pamela (1999). Is social capital declining in the United States? A multiple indicator assessment. *American Journal of Sociology*, 105: 88–127.

Peters, Bernhard (2002). A new look at "national identity". *Archives of European Sociology*, 43: 3–32.

Peterson, Steven A. and Maiden, Robert J. (1993). *The Pubic Lives of Rural Older Americans*. Lanham, MD: University Press of America.

Phelan, Jo A., Link, Bruce G., Stueve, Ann and Pescosolido, Bernice A. (2000). Public conceptions of mental illness in 1950 and 1999: What is mental illness and is it to be feared. *Journal of Health and Social Behavior*, 41: 188–207.

Popenoe, David and Whitehead, Barbara DeFoe (1999). The state of our unions. Report of the National Marriage Project, Rutgers University.

Presser, Stanley (1984). The use of survey data in basic research in the social sciences. In Charles F. Turner and Elizabeth Martin (eds), *Surveying Subjective Phenomena*, Vol. 2. New York: Russell Sage Foundation.

Putnam, Robert D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Ramirez, Carl (1990). Response effects and mode of administration or: More reasons why the GSS should remain an in-person survey. Unpublished paper.

Rindfuss, Ronald R., Brewster, Karin L. and Kavee, Andrew L. (1996). Women, work, and children: Behavioral and attitudinal change in the United States. *Population and Development Review*, 22: 457–482.

Schuman, Howard and Krysan, Maria (1997). A historical note on White beliefs about racial inequality. *American Sociological Review*, 64: 847–855.

Schuman, Howard, Steeh, Charlotte, Bobo, Lawrence and Krysan, Maria (1997). *Racial Attitudes in America: Trends and Interpretations*, 2nd edn. Cambridge: Harvard University Press.

Smith, Tom W. (1994). Is there real opinion change? *International Journal of Public Opinion Research*, 6: 187–203.

Smith, Tom W. (1997). Liberal and conservative trends in the United States since World War II. *Public Opinion Quarterly*, 54: 479–507.

Smith, Tom W. (1997). Societal change in America, 1972–1996: Time trends in the National Opinion Research Center's General Social Survey. *SINET*, 49: 1–3.

Smith, Tom W. (1999). The emerging 21st century American family. *GSS Social Change Report No. 42*. Chicago: NORC.

Smith, Tom W. (2000). Does knowledge of science breed confidence in science? In Jon D. Miller (ed.), *Perceptions of Biotechnology: Public Understanding and Attitudes*. New York: Hampton Press.

Smith, Tom W. (2005a). *Jewish Distinctiveness in America*. New York: American Jewish Committee.

Smith, Tom W. (2005b). The laws of studying societal change. *Survey Research*, 26: 1–5.

Smith, Tom W. and Jarkko, Lars (1998). National pride: A cross-national analysis. *GSS Cross-National Report No. 19*, Chicago: NORC.

Smith, Tom W. and Kim, Seokho (2006). National pride in comparative perspective: 1993/95 and 2003/04. *International Journal of Public Opinion Research*, 18: 127–136.

Smith, Tom W., Koch, Achim, Park, Alison and Kim, Jibum (2006). Social-science research and the general social surveys. *Comparative Sociology*, 4, forthcoming.

Smith, Tom W. and Stephenson, C. Bruce (1979). An analysis of test/retest experiments on the 1972, 1973, 1974, and 1978 general social surveys. *GSS Methodological Report No. 8*, Chicago: NORC.

Straus, Murray A. and Mathur, Anita K. (1996). Social changes and trends in corporal punishment by parents from 1968 to 1994. In Deltlev Freshsee, et al. (eds), *Family Violence Against Children: A Challenge to Society*. New York: Walter de Gruyter.

Verba, Sidney and Nie, Norman H. (1972). *Participation in America: Political Democracy and Social Equality*. New York: Harper and Row.

Walker, Melissa E. (1997). A longitudinal analysis of teacher values from 1973 to 1994. Unpublished Ph.D. dissertation, Texas A&M University.

Weaver, Charles N. and Trankina, Michele L. (1996). Should scientists and engineers be combined into one analysis sample? *Psychological Reports*, 79: 1151–1153.

Chapter 4

Structuring the National Crime Victim Survey for use in longitudinal analysis

Lawrence Hotchkiss and Ronet Bachman

1 Introduction[1]

There are many research questions that can only be answered with data that follow individuals over time. Research in the area of crime victimization is no exception. In fact, a burgeoning area of research in criminology is related to issues of recurring victimization both at the individual level (Farrell, Tseloni and Pease, 2005; Lauritsen and Quinet, 1995; Menard, 2000; Stevens, Ruggiero, Kilpatrick, Resnick and Saunders 2005) and the aggregate household or neighborhood level as well (Bowers and Johnson, 2005; Farrell, Sousa and Weisel, 2002; Outlaw, Ruback and Britt, 2002). Although the National Crime Victimization Survey (NCVS) is most often utilized in a cross-sectional format or to estimate aggregate trends, it has the capacity for individual-level longitudinal analysis, because each sample unit (address) stays in the sample for three-and-one-half years. Despite the potential to reconfigure NCVS data files longitudinally, very few have undertaken the task, primarily because of the many challenges one encounters when doing so. Issues of attrition plague most surveys that attempt to track individuals over time, however the issue is even more complex for the NCVS. For example, the NCVS samples residential addresses, not households or individuals, so different household members may move in and out of an address, or a different household may move in altogether. Another stumbling block is related to multiple victimizations, particularly those that occur within the same month. The only information the NCVS collects about the time ordering in a given interview is the month of occurrence. Moreover, if a respondent has experienced six or more similar victimizations and can't recall them separately, these incidents are recorded as one series crime. The only information the NCVS collects about the time ordering of series crimes is the number of incidents per quarter. These, along with other challenges, confront researchers eager to exploit the longitudinal structure of the NCVS.

2 NCVS procedures

The NCVS is an ongoing survey of personal and household victimizations designed to be representative of all persons living in noninstitutional households in the United States over

[1] The authors wish to thank Jeremy Shimer of the US Census Bureau and Thomas Zelenock of the ICPSR for valuable reviews and suggestions related to this chapter.

12 years of age. Begun in 1973, the NCVS was designed with four primary objectives: (1) to develop detailed information about victims and consequences of crime, (2) to estimate the number and types of crimes not reported to the police, (3) to provide uniform measures of selected types of crimes, and (4) to permit comparisons over time and types of areas. The survey categorizes crimes as "personal" or "property." Personal crimes include rape and sexual attack, robbery, aggravated and simple assault, and purse-snatching/pocket-picking, while property crimes include burglary, theft, motor vehicle theft, and vandalism. Respondents are asked a series of questions designed to determine whether she or he was victimized during the six-month period preceding the first day of the month of the interview. A "household respondent" is also asked to describe crimes against the household as a whole (e.g., burglary, motor vehicle theft). An incident report for each victimization includes information such as type of crime, month, time of day, and location of the crime, information about the offender including the number of offenders, perceived gang membership, their race, gender, and age, and the relationship between victim and offender, self-protective actions taken by the victim during the incident and results of those actions, consequences of the victimization including injuries sustained, type of property lost, whether the crime was reported to police and reasons for reporting or not reporting, and offender use of weapons, drugs, and alcohol. The NCVS also collects the respondent's demographic information such as age, race, gender, income, and occupation.

The NCVS uses a stratified, multistage cluster sampling design. The primary sampling units (PSUs) are counties, groups of smaller counties, and metropolitan areas. A sample of census-identified enumeration districts is selected from PSUs; these districts are geographic areas that encompass approximately 750–1500 persons and range in size from a block to hundreds of square miles. Enumeration districts are divided into clusters of approximately four housing units. The final sampling procedure randomly selects clusters of housing units. The resulting sample consists of approximately 43,000 housing units and other living quarters. These housing units remain in the sample for a period of three-and-a-half years. Every six months, a new group of housing units replaces one-seventh of the housing units then in the sample (For a more detailed discussion, see the NCVS codebook, available from ICPSR or the Bureau of Justice Statistics, US Department of Justice, 2006).

3 Individual-level longitudinal analyses using the NCVS

There are numerous research questions that could be investigated using these data in a longitudinal form. We will highlight a few studies that have successfully done so, each investigating different research questions. One of the first successful attempts to merge the individual files in the NCVS into a longitudinal format was published by Mark Conaway and Sharon Lohr (1994). They were interested in examining the factors related to the police reporting behavior of crime victims. Unlike most of the extant research on this topic that primarily utilizes information about the incident, such as crime seriousness, Conaway and Lohr were interested in the effects of previous reporting behavior on future reporting behavior. Specifically, they investigated whether victims who reported positive experiences with the police in a previous interview would be more likely to report victimization again than respondents who either had negative experiences with reporting to the police in the past, or had never reported a victimization to police at all. After controlling for other characteristics of the crime and demographics of the respondent, they found that victims were more likely to report a violent crime to police if a previous victimization had been reported and if police did routine follow-up activity or if an arrest was made or property was recovered.

Most recently, researchers using NCVS data from 1998 to 2000 found that investigatory efforts by police in a prior victimization increased the likelihood of victims reporting an ensuing victimization only when the victim, rather than someone else, reported the prior victimization (Xie, Pogarsky, Lynch, and McDowall, 2006). Xie and colleagues also found that an arrest after an individual was victimized in the past had no effect on whether the individual reported an ensuing victimization (2006).

Repeat victimization has become a major focus of criminological researchers for several reasons. Perhaps foremost of these is the fact that offending behavior, victimizations, and offending locations each tend to cluster (Farrell, Tseloni, and Pease, 2005). Estimating the factors related to repeat victimization, then, can also help illuminate patterns of offending and “hot spots” for crime. Information regarding repeat victimization, then, can inform crime control policies as well as the development of criminological theory (Laycock, 2001). Repeat victimization has been found to account for a significant proportion of victimization in juveniles (Lauritsen and Davis Quinet, 1995; Menard, 2000), and for adults (Farrell, Tseloni, and Pease, 2005; Gabor and Mata, 2004). After examining the rates of repeat victimization in the NCVS, Graham Farrell and his colleagues (2005) concluded that repeat victimization accounted for about one-quarter of all assaults and sexual assaults, about 20% of robberies and other personal thefts, and about 18% of burglaries.

Lynn Ybarra and Sharon Lohr (2002) examined estimates of repeat victimization using the NCVS and the effects of repeat victimization on attrition, particularly the effects of being victimized by an intimate partner. Importantly, what they found was that victims of intimate partner violence were more likely to drop out of the survey than nonvictims. This type of research has important implications for magnitude estimates of victimization in general, and for intimate partner violence in particular. If individuals who drop out of the survey (e.g., move or refuse to respond) are more likely to be repeat victims than are individuals who replace them, then cross-sectional estimates of victimization may be biased downwards.

Catherine Gallagher (2005) examined the extent to which injured crime victims who sought medical care for their injuries were more or less likely to sustain injuries as the result of a future victimization. Specifically, she compared the probability of injury recurrence for injury victims of crime who sought medical help for their injuries to the probability for those who had not sought medical help. Gallagher (2005), in fact, found that victims who received medical care for earlier injuries resulting from violence were at a decreased risk of sustaining a future violence-related injury even after controlling for the seriousness of the injury and medical coverage. Gallagher also found that past medical care more generally protected individuals from future violence. Although she was not able to isolate the mechanism for this finding, it does appear that something about medical treatment has a violence-reducing effect. Gallagher concludes that efforts should be made to secure medical treatment for victims of crime in general.

3.1 Structure of the NCVS full hierarchical data files

The National Crime Survey (NCS) was an ongoing longitudinal crime-victimization survey of households and persons. Begun in 1973, it underwent a major revision in the years leading up to 1992 when it was renamed the National Crime Victimization Survey (NCVS). This chapter describes data associated with the full hierarchical files of the NCVS. The NCVS selects a new sample every three years, numbered 15 through 22 to date. Earlier sample numbers are associated with the NCS.

Addresses, not individuals nor households, are the units selected into each sample. Residents of selected addresses are scheduled to be interviewed seven times. Questions about victimization incidents refer to the six months prior to the interview, called the reference period. The first interview is used for "bounding" victimizations within the reference period to eliminate over-reporting errors due to telescoping. Telescoping refers to the tendency of respondents to report crimes that occurred before the reference period. Victimization incidents reported during the first interview are not used in the estimation of victimization rates. Instead, they are used to identify incidents reported during interview two that occurred before the six-month reference period. Data from the first interview are excluded from the public-use data files. There is, however, one exception to this. If a new household moves into a sample address, no bounding interview is conducted for this replacement household, and its first interview is included in the public-use data. Hence, household/person interview numbers for nonreplacement households span the interval 2 through 7, inclusive, but interview numbers for replacement households vary over the interval 1 through 6, inclusive.

Each NCVS sample is divided into six rotation groups, numbered 1 through 6, and each rotation group is divided into six panels, numbered 1 through 6. Interview 1 of rotation-group 1, panels 1 through 6 occurs during January through June, respectively of the year the sample is activated. The first interview of rotation-group 2, panels 1 through 6 occurs each July through December of the same year. The second interview of rotation group 1, panels 1 through 6 occurs during July through December of year 1. Hence the second interview of rotation-group 1 occurs at the same time as the first interview of rotation-group 2.

This staggered design extends to all six rotation groups and panels within rotation groups. Since each address is interviewed every six months and stays in the sample for seven interviews, beginning with the last interview (7) of the first sample, seven rotation groups are interviewed simultaneously, some from one sample and some from the next numbered sample in the sequence. But interviews from just six rotation groups appear in the data for a given date, because the bounding interviews are not included in the public-use files, and replacement households can be interviewed a maximum of six times.

Table 4.1 summarizes this structure for the data contained in the 2003 full-sample file. This file contains data for all of the interviews in 2003 and the first six months of 2004. Sample designations are given as header rows, prefixed in the documentation with the letter "J" to designate a survey done by the US Census Bureau for the US Department of Justice.

The address-interview number appears on the line labeled "ADier_no". This is the same value as what the Census Bureau calls time in sample (TIS). The main cell entries give the panel (tens digit) and rotation numbers (ones digit). The row labels indicate the year and month in which residents of the address are scheduled to be interviewed. Empty cells indicate that no panel/rotation combination in the specified sample/rotation group/panel is scheduled to be interviewed during the year and month given by the row label. For example, the empty cell in the first column and row "2003 Jan" indicates all seven interviews were completed before January 2003 for respondents in sample J20, panel 1, rotation-group 1. The empty cell in the last column of the same row ("2003 Jan") indicates that interviewing had not begun by January 2003 for sample J21, panel 1, rotation-group 6. The first interview for this group (sample J21, panel 1, rotation-group 6) began in July 2003, as indicated by the "16" (panel 1, rotation 6) entry in the last column

Table 4.1 Rotation chart for selected years, samples J20 and J21[a]

	Sample number											
Year and month	*J20*						*J21*					
ADier_no:					7	6	5	4	3	2	1	
2003 Jan					15	16	11	12	13	14	15	
Feb					25	26	21	22	23	24	25	
Mar					35	36	31	32	33	34	35	
Apr					45	46	41	42	43	44	45	
May					55	56	51	52	53	54	55	
Jun					65	66	61	62	63	64	65	
ADier_no:						7	6	5	4	3	2	1
2003 Jul						16	11	12	13	14	15	16
Aug						26	21	22	23	24	25	26
Sep						36	31	32	33	34	35	36
Oct						46	41	42	43	44	45	46
Nov						56	51	52	53	54	55	56
Dec						66	61	62	63	64	65	66
Year and month	*J20*						*J21*					
ADier_no:							7	6	5	4	3	2
2004 Jan							11	12	13	14	15	16
Feb							21	22	23	24	25	26
Mar							31	32	33	34	35	36
Apr							41	42	43	44	45	46
May							51	52	53	54	55	56
Jun							61	62	63	64	65	66

[a] Extrapolated and adapted from "NCS/NCVS Rotation Chart:: July 1994–June 1998" (US Department of Justice, Bureau of Justice Statistics, 2006: pp. 443–444)

of row "2003 Jul." Similarly, the "34" in column 10, of the row for April 2004 indicates that sample J21, panel 3, rotation-group 4 was scheduled for its fourth interview during April of 2004 (read the interview number from the header rows titled "ADier_no").

The full NCVS data collection is supplied by the Inter-university Consortium for Social and Political Research (ICPSR) at the University of Michigan. The full data are stored in ASCII format using a hierarchical structure. There are four types of records in the full data files:

- Address: One record per address per completed contact.
- Household: One record per household per completed contact.

- Person: One record per person per completed household interview.
- Incident: One record per person/household per completed interview per crime-victimization incident.

All four record types are supplied in a single file containing a record-indicator flag.

A contact is not the same as a completed interview. A variable in the household data indicates whether each contact was, in fact, a completed interview. For the data files to date, over eighty percent (80.5%) of the household records indicate completed household-level interviews for all eight $(22 - 15 + 1 = 8)$ NCVS samples combined. Since a contact does not occur for every scheduled interview, the number of records per address, household, and person varies, with a maximum of six in public-release data. The number of incident records per person varies from zero to the number of victimization incidents reported to have occurred during the reference period (prior six months). The exception to this is for series incidents, which are recorded on just one incident record. An incident is classified as a series incident if an individual experiences six or more incidents within a reference period and can't recall enough details of each to report them as separate incidents. There is, however, no theoretical upper limit on the number of incident records.

3.2 Restructuring the NCVS data

The hierarchical organization of the NCVS data minimizes the space needed to store it. A "rectangular file" containing all the data would be excessively large. Nonetheless, most statistical packages are not designed to analyze directly data stored in a hierarchical format like that of the NCVS.

Tables 4.2, 4.3 and 4.4 present a highly simplified example of the restructuring needed to produce a rectangular person-level file containing gender (V3018, constant over interviews) and one variable for each of the seven interviews indicating (1) household income, (2) age, and (3) marital status at the time of the interview. Additionally, two variables per interview allow for up to two crime-victimization reports per interview.

The layout for these variables for two households and three persons in the NCVS hierarchical format appears in Table 4.2. Where feasible, the table contains the variable names used in the NCVS documentation (e.g., V2026 for household income). But the household ID (HH_ID), person ID (pers_ID) and household interview number (HHier_no) must be constructed from more than one NCVS variable and are given names as shown in Table 4.2. For clarity, the table displays the three types of records in separate blocks, but the NCVS hierarchical data files are sorted by address ID and record type within the address ID.[2]

Table 4.3 shows the households, individuals, and victimization incidents merged together by household ID (HH_ID), person ID (pers_ID), and household interview number (HHier_no). This produces a file with data from each interview appearing on separate records. This format sometimes is called "long" format.

Table 4.3 illustrates how this merge generates (1) duplication of data, and (2) a large amount of missing data. The duplication arises from two sources: repeating household information on each person record and repeating household and person information for each incident report. In Table 4.3, all the household information for household 1 is listed twice, since two persons in the example reside in household 1, and both household and person variables are replicated for household 2, person 1,

[2] A complete description of how to use the identification variables to construct a working rectangular file from the NCVS hierarchical files may be obtained directly from the first author (Lawrence Hotchkiss, larryh@udel.edu or at the following URL: http://gorilla.us.udel.edu/ncvs//NCVSdataPreparation.doc)

Table 4.2 Hierarchical structure of the NCVS full data files

Household Records		
HH_ID	*HHier_no*	*V2026*
1	2	13
1	3	13
1	4	13
1	5	13
1	6	14
1	7	14
2	3	9
2	4	98
2	5	8
2	7	9

Person Records					
Pers_ID	*HH_ID*	*HHier_no*	*V3014*	*V3015*	*V3018*
1	1	2	47	1	1
1	1	3	47	1	1
1	1	4	48	1	1
1	1	5	48	1	1
1	1	6	49	4	1
1	1	7	49	4	1
2	1	2	15	5	2
2	1	3	16	5	2
2	1	4	16	5	2
2	1	5	17	5	2
2	1	6	17	5	2
2	1	7	18	5	2
1	2	3	35	4	2
1	2	4	35	4	2
1	2	5	36	4	2
1	2	7	37	4	2

Incident Records			
Pers_ID	*HH_ID*	*HHier_no*	*V4529*
1	2	5	17
1	2	5	20

Definitions:
1 HH_ID = Household ID, constructed from 3 NCVS variables
2 HHier_no = Household interview number, calculated from NCVS variables (see below)
3 V2026 = Household income: 8: $20,000 ≤ inc < 25,000; 9: $25,000 ≤ inc < 30,000; 13: $50,000 ≤ inc < 75,000; 14: inc ≥ $75,0000
4 Pers_ID = Person ID, constructed from 4 NCVS variables
5 V3014 = Age (years to last birthday, allocated)
6 V3015 = Marital status (current interview): 1 = Married; 2 = Widowed; 3 = Divorced; 4 = Separated; 5 = Never married; 8 = Residue; 9 = Out of universe
7 V3018 = Gender (1=male; 2=female, allocated; 8 = Residue; 9 = Out of universe)
8 V4529 = Type of crime (TOC) classification: 17 = Assault without weapon without injury, 20 = Verbal threat of assault

Table 4.3 Merged household, person and incident records by HH_ID, Pers_ID and ADier_no (long format)

HH_ID	*Pers_ID*	*HHier_no*	*V2026*	*V3014*	*V3015*	*V3018*	*V4529*
1	1	2	13	47	1	1	.
1	1	3	13	47	1	1	.
1	1	4	13	48	1	1	.
1	1	5	13	48	1	1	.
1	1	6	14	49	4	1	.
1	1	7	14	49	4	1	.
1	2	2	13	15	5	2	.
1	2	3	13	16	5	2	.
1	2	4	13	16	5	2	.
1	2	5	13	17	5	2	.
1	2	6	14	17	5	2	.
1	2	7	14	18	5	2	.
2	1	3	9	35	4	2	.
2	1	4	98	35	4	2	.
2	1	5	8	36	4	2	17
2	1	5	8	36	4	2	20
2	1	7	9	37	4	2	.

Note: See Table 4.2 for definition of the variables.

interview 5. The missing data are due to the fact that most households and persons report no victimizations during each interview. In Table 4.3, for example, 15 out of the 17 values for type of crime (V4529) are missing.

Table 4.4 summarizes a layout variously called a rectangular file, flat file or wide-format file. The data in this example produce 39 variables in the rectangular file, in contrast to just 8 in the long-format file in Table 4.3. The conventional representation of a data file associates each variable with a column and each person (observation) with a row, but a table containing 39 columns and three rows cannot be displayed on the printed page. Table 4.4 shows the transposition of the usual representation of a data file with the transposed observations wrapped into three columns to conserve space.

The rectangular file represents data for the different interviews and incidents by different variables (columns rather than rows). Gender (V3018) appears just once per person, since it does not change. A variable must be defined for each of the other non-ID household and person variables for each of the seven possible interview numbers, because they are not necessarily constant over interviews. In the example, each of these new variables is given a numeric suffix corresponding to the interview number, e.g., V3015_1, V3015_2, . . . , V3015_7.

Variables appearing in the household records are attached to each person in the household, generating substantial duplication. Variables appearing on the incident records generate even more variables in the rectangular file. A variable must be defined for each interview and each incident per interview. In the example, the number of incidents is capped at two, so 14 variables are needed to represent type of crime, which requires just one variable in the incident records. In Table 4.4, variable names for the incident variables are formed by adding a double suffix to the original variable name, a letter to index incidents and a number to index interviews (V4529_A1 . . . V4529_B7).

Even though no person or household appears in the data for more than six interviews, seven variables must be reserved for each variable in

Table 4.4 “Flat file” merged household, person and incident records

No.	*Variable*	*Values*		
1	HH_ID	1	1	2
2	Pers_ID	1	2	1
3	V2026_1	.	.	.
4	V2026_2	13	13	.
5	V2026_3	13	13	9
6	V2026_4	13	13	98
7	V2026_5	13	13	8
8	V2026_6	14	14	.
9	V2026_7	14	14	9
10	V3014_1	.	.	.
11	V3014_2	47	15	.
12	V3014_3	47	16	35
13	V3014_4	48	16	35
14	V3014_5	48	17	36
15	V3014_6	49	17	.
16	V3014_7	49	18	37
17	V3015_1	.	.	.
18	V3015_2	1	5	.
19	V3015_3	1	5	4
20	V3015_4	1	5	4
21	V3015_5	1	5	4
22	V3015_6	4	5	.
23	V3015_7	4	5	4
25	V3018	1	2	2
26	V4529_A1	.	.	.
27	V4529_A2	.	.	.
28	V4529_A3	.	.	.
29	V4529_A4	.	.	.
30	V4529_A5	.	.	17
31	V4529_A6	.	.	.
32	V4529_A7	.	.	.
33	V4529_B1	.	.	.
34	V4529_B2	.	.	.
35	V4529_B3	.	.	.
36	V4529_B4	.	.	.
37	V4529_B5	.	.	20
38	V4529_B6	.	.	.
39	V4529_B7	.	.	.

the NCVS household and person data that is not constant over all interviews. And seven times the number of incidents is required for variables residing on incident records. Seven interview numbers are needed, even though no individual or household appears in the data more than six times, because interview numbers for replacement households span 1 to 6, inclusive, but interview numbers for nonreplacement households span 2 to 7, inclusive.

As you can see, the wide-format file generates enormous redundancy. The average size of households is about 2.2, so household variables are repeated over two times on average in a file where observations are individuals. Table 4.5 gives additional indication of the magnitude of redundancy. It shows a frequency distribution of the number of interviews and a cross-tabulation of number of incidents by household interview number. The top panel of the table shows, for example, that 255,854 out of a total of 737,946 (34.7%) persons completed just one interview. For these persons, six of their seven household and person variables in each sequence would be missing in a rectangular file. The bottom panel of Table 4.5 indicates an even more extreme proportion of missing data. When provision is made for the maximum number of incidents (13), each variable on the NCVS incident records generates $7 \times 13 = 91$ variables in the rectangular file, V4529_A1 . . . V4529_M7, for instance. Yet, as the tabulation demonstrates, the vast majority of these values are missing.[3]

Often, however, one needs to keep just one variable per interview in the wide-format file for each variable in the incident records. Many analyses require only summaries of incidents, such as the number of violent-crime victimizations reported per interview, or simply a flag set to 1 if one or more violent victimizations is reported, and zero otherwise. In these situations, each variable used from the incident file requires just one variable per interview rather than 13.

[3] The 782,316 persons in our working data file generate $91 \times 737{,}946 = 67{,}153{,}086$ data cells for each variable in the incident records kept in the flat file. Of these, 66,989,098 contain missing values (99.8%, assuming incident records contain no missing values).

Table 4.5 Distribution of number of interviews and number of incidents by interview number

Number of interviews	*Frequency*	*Cumulative frequency*
1	255,854	255,854
2	150,825	406,679
3	101,584	508,263
4	76,836	585,099
5	64,282	649,381
6	88,565	737,946

Cross-tabulation: Household interview number (HHier_no) by number of incidents

Number of incidents	*Interview number*							
	1	2	3	4	5	6	7	*Total*
0	151,712	375,746	323,818	289,046	265,803	246,185	235,684	1,887,994
1	17,222	26,175	19,204	15,346	12,246	11,715	10,928	112,836
2	3132	4140	2662	2044	1446	1413	1348	16,185
3	794	984	607	412	272	314	287	3670
4	275	322	182	99	90	90	86	1144
5	88	102	59	41	25	31	30	376
6	22	36	24	14	7	14	8	125
7	11	9	4	6	3	2	1	36
8	4	3	2	1	0	1	0	11
9	2	2	2	2	1	1	2	12
10	2	2	1	0	1	1	0	7
11	0	2	0	0	0	0	0	2
13	0	1	0	0	0	1	0	2
Total	**173,264**	**407,524**	**346,565**	**307,011**	**279,894**	**259,768**	**248,374**	**2,022,400**

Notes: (1) Cell entries in the top panel are counts of persons, and cell entries in the bottom panel are person records, both accumulated over all NCVS samples, 15 through 22, excluding cases with out-of-range address numbers and inconsistent reports of gender; (2) Excludes type Z noninterviews.

Analyses reported later in this chapter compare a model reported by Janet Lauritsen (2001) for all interviews combined to models with the same specification broken out by interview number. These analyses were done twice, once using a wide-format file and once using the long-format file. In the first instance, the analysis for each interview number used variables such as age2, age3, . . . , age7, and the second approach used "by-variable" processing. All the numeric output from the two approaches match exactly. The long-format file contains a subset consisting of 140 variables. A wide-format file capping the number of incidents at 5 and including just 72 of the variables in the long-format file generated 885 variables. The size of the long-format file is just over one gigabyte, and the size of the wide-format file is over two gigabytes (setting the default variable length = 3 in both cases).

3.3 Attrition and nonresponse

Even with the high response rate (95%) reported in the NCVS documentation, only about 23% of addresses appear in the household records for all six possible interviews in the public-use files, and over 15% were, in fact, classified as interviewed all six times. This is due in part to the fact that the maximum number of possible interviews is less than seven for some households, but most of it is due to attrition, and much attrition occurs at the household and person levels.

Nonetheless, the attrition rates are good by longitudinal survey standards. As Table 4.6 reports, about 70% of persons interviewed during interviews 2 through 6 also were interviewed during their next scheduled interview. But the table also shows that the percentage of persons who are interviewed after one or more intervening interviews rapidly dwindles. For example, of those interviewed at interview 2, 53.14% were interviewed at interview 4, and this declines to 27.52% by interview 7. These percentages understate attrition, since not all of those completing interview 2 and interview 7, for example, completed all the intervening interviews.

Additionally, not all person records in the NCVS data files are associated with completed interviews. If a household is classified as a completed interview but an individual member of the household could not be interviewed, the noninterview is classified as a Type Z noninterview, and the person record for this person appears in the public-use data with personal information carried forward from a previous interview. The person weight (V3080) for type Z records is set to zero, and a zero weight appears to be the only definite indication of a noninterview. The interviewed flags in Table 4.6 were set to zero when person weights were zero.[4]

Table 4.6 Percentage of persons interviewed during an interview who also were interviewed during the subsequent interview

Subsequent interview number	*First interview number*				
	2	3	4	5	6
3	68.59				
4	53.14	69.02			
5	41.95	54.53	70.65		
6	33.84	44.14	57.48	72.47	
7	27.52	36.27	47.66	60.24	74.42

Note: Unweighted estimates.

4 Example: Prediction of violent crime victimization

One of the primary motivations for reformatting the NCVS data into a long-format or wide-format file is to test models of the determinants of crime victimization. An excellent example of this type of work is contained in a paper by Janet Lauritsen (2001). She reports several logistic-regression models predicting violent-crime victimization using as regressors: age, gender, marital status (married), household income, length of residence at the current address, residence in the central city of an MSA, and three contextual-level variables defined by features of the census track containing sample address—a disadvantagement index, an immigration index and an index of area instability. One set of six models uses only the individual-level predictors. One of these six models applies to all violent victimizations in all geographic areas. A second model is restricted to incidents reported to have occurred in the neighborhood of one's residence. The other four are for: non-central-city and all neighborhoods; non-central-city within neighborhood; central-city and all neighborhoods; and central-city within neighborhood. A second set of analyses adds the contextual-level variables to the set of individual predictors.

[4] See the NCVS Codebook (US Department of Justice, 2006, p. 419). The authors are grateful to Jeremy Shimer of the US Census Bureau for clarifying this issue.

The paper uses a special release of the NCVS data for 1995 containing confidential identifiers needed to match the NCVS sample addresses to census tracts as required to merge the contextual variables onto NCVS data. NCVS data from households, persons, and incidents were merged into a single long-format analysis file containing one record per person per incident. Inferential statistics were corrected for the dependence among the observations.

This analysis strategy that implies the model is correct for all interviews, including the first unbounded interview for replacement households. It also precludes a precise estimate of the risk of victimization during a six-month interval, because the dependent variable does not exactly indicate whether an individual was a victim of at least one violent-crime incident during the reference period. Instead, the dependent variable for each person is assigned 0 if no incident records match to the person and 1 for each matching incident record containing a type-of-crime code defined as a violent crime. Since few persons report more than one incident per interview (Table 4.5), however, this latter consideration probably has little practical impact.

This chapter extends the work reported by Lauritsen (2001) by (1) estimating separate logistic-regression models for each interview, 2 through 7, and (2) defining the dependent variable to be an indicator set equal to 1 for persons who reported at least one victimization classified as a violent crime during each interview and 0 otherwise.[5] Predicted probabilities from a binary regression with this measure of victimization therefore indicate the risk of being the victim of a violent crime during a six-month interval.

We report comparisons to the Lauritsen model for overall violent-crime victimization using individual-level regressors. The variables are:

VCvictim (dependent variable)	Violent-crime victim: Victim of at least 1 violent crime during the reference period for each interview [at least one incident record associated with each person had type-of-crime code (v4529) classified in the interval 1 through 14, inclusive].
Age	Age in years (v3014)
Male	Dummy variable for male: male = 1 (v3018 = 1)
Black	Dummy variable for race reported as Black (v3023 = 2)[6]
Married	Dummy variable, married = 1 (v3015 = 1)
HHinc	Household income (v2026 in original 14 ordered categories)
eveningsIn	Ordinal indicator of frequency of spending evenings at home (v3029)
Tenure	Number of years lived at current residence (constructed from v3031 and v3032)
cntrCityMSA	Lived in the central city of an MSA (v2129 = 1)
IncomeMDD	Missing-data dummy for HHinc
EveningsMDD	Missing-data dummy for eveningsIn
TenureMDD	Missing-data dummy for tenure

[5] We used Lauritsen's definition of the dependent variable for our replication of her model, column 1, Table 4.7)

[6] Year-file 2002 and later used the expanded census race categories permitting respondents to select mixtures of races. The Black dummy variable was coded to 1 if any mixture including Black was checked. The analogous definition was applied to White (used in the subsetting for all years, column 8 of Table 4.7).

Although Lauritsen (2001) did not mention missing-data dummy variables, we had to eliminate missing values for income, evenings spent in house, and tenure to match (approximately) the sample size she reports. We set missing values for these three variables to their respective means and set the companion missing-data-dummies to 1. The coefficients associated with the missing-data-dummies therefore estimate the degree to which the predicted probability of being a victim of violent crime for those with missing values differs from the predicted probability using the mean on the corresponding "parent" variable.

Table 4.7 reports (1) our replication of the Lauritsen model for overall violent-crime victimization (column 1), (2) a set of analyses consisting of one model for each of the (potentially) bounded interviews, 2 through 7 with the base year restricted to 1994, 1995 and 1996 (columns 2 through 7), and (3) one model for interview 2 with no restriction on the base year and year included as one of the regressors (last column). The restrictions on the base year were imposed to confine the analysis to the time period around 1995. All the calculations exclude cases of inconsistent gender reports and persons not classified as either Black or White. The latter restriction follows the subsetting used by Lauritsen. The replication of Lauritsen's model also excludes person records with invalid interview numbers, and the rest of the models exclude individuals residing in households for which one or more address-interview numbers are out of range (because valid household/person interview numbers cannot be assured unless all address-interview numbers are valid for a household or person).

The replication sample (column 1) contains 173,487 person/incident records, after eliminating all persons with invalid address IDs and persons not classified as Black or White. Lauritsen reports 171,949 observations for her individual-level sample. So our replication sample is slightly larger than the Lauritsen sample, and we have not been able to determine the reason(s). Column 1 shows estimates close to those reported in Lauritsen's published work. The parameter estimates and pattern of statistical significance also match closely, except that we get a coefficient for Black of 0.202 instead of 0.021 reported by Lauritsen.

The six models designed to check whether effect estimates depend on the interview number (columns 2 through 7) conforms roughly to the Lauritsen results; there certainly are no dramatic departures from her estimates. But the sample sizes are large enough that it is unlikely the variation over interviews is entirely due to sampling error. And there is enough variation across the interviews to suggest additional work is needed to determine the cause of it. If attrition were random, model estimation would not depend on the interview number. Yet, the estimated coefficient for male at interview 2 (OR = 1.51) is about 3.4 times the estimate at interview 7 (OR = 1.13). In fact, the absolute magnitude of every effect estimate, except tenure, central-city residence and the missing-data-dummy variables for interview 7, is smaller than the corresponding magnitude for interview 2. The estimated effect of Black varies quite substantially over the interviews. Even with the large sample, it more often is not significant than significant. Lauritsen also found variation in the effect of Black among the several models she reported. The effect of age remains fairly consistent until interview 5 when it begins to decline, ending at interview 7 at just under half its earlier magnitudes. The estimated effect of tenure (length of residence at the current address) also varies sporadically over interviews.

The last column of Table 4.7 reports estimates of the model with no restriction on the base year and the variable year added as one predictor. These estimates also conform fairly well to the Lauritsen model but, again, indicate several modest deviations from it. The most interesting

Table 4.7 Logistic regressions predicting violent-crime victimization

	1994 ≤ Base Year ≤ 1996							*All years*
Parameter	*Lauritsen[1] estimate*	*Interview 2 estimate*	*Interview 3 estimate*	*Interview 4 estimate*	*Interview 5 estimate*	*Interview 6 estimate*	*Interview 7 estimate*	*Interview 2 estimate*
Intercept	−2.541***	−2.820***	−2.816***	−3.226***	−3.045***	−3.386***	−4.043***	71.448***
Year								−0.037***
Age	−0.031***	−0.033***	−0.032***	−0.031***	−0.025***	−0.025***	−0.015***	−0.027***
Male	0.334***	0.410***	0.326***	0.404***	0.390**	0.288*	0.120	0.360***
Black	0.202*	0.188+	0.107	−0.020	0.136	−0.049	0.138	0.115+
Married	−0.694***	−0.682***	−0.590***	−0.759***	−0.680***	−0.555***	−0.478***	−0.708***
Income	−0.047***	−0.032***	−0.041**	−0.029+	−0.059**	−0.019	0.006	−0.033***
Evenings in	−0.098**	−0.066+	−0.080	−0.048	−0.093	−0.122+	−0.023	−0.086***
Tenure	−0.037***	−0.012*	−0.019*	−0.013+	−0.033**	−0.015	−0.020*	−0.015***
Central city	0.322***	0.314***	0.288**	0.506***	0.237+	0.345*	0.388**	0.270***
MDD income	−0.299*	−0.103	−0.161	−0.320+	−0.017	−0.266	−0.324	−0.194**
MDD evenings home	−0.037	−0.274	−0.613	−1.042	−0.044	−1.126	−1.254	−0.329***
MDD central city	−1.143+	−0.382	−0.174	−0.087	−0.028	0.699	−1.035	−0.596+
Sample size	173,487	100,213	83,237	71,820	64,815	58,948	54,686	387,368

[1] All interviews, 1995

*** p ≤ 0.0001 ** p < 0.001 * p ≤ 0.01 + p ≤ 0.05

finding here is the strong negative effect of year. The odds ratio for the entire 13 years covered by these samples is

$$\text{OR} = \exp(-13^{*}0.037) = \exp(-0.481) = 0.618.$$

That is, the odds of being a victim of a violent crime have declined by a factor of 0.618 over the 13-year period. This finding agrees with estimates published by the Bureau of Justice Statistics that show a steady decline in the incidence of violent crime since 1993 (BJS, 2006). It is also noteworthy that the negative effect of year occurs even with a control for age, suggesting that the decline in the rate of reported violent-crime incidents is not due entirely to aging of the population.

It is important to note that attrition in the sample is not random. Table 4.8 shows results of a logistic regression model predicting nonattrition from interview 2 to interview 3 ($1994 \leq$ base year ≤ 1996).

The table indicates that attrition is dependent on all the variables in the Lauritsen model, with very small p-values. The small p-values, however, reflect the very large sample size. The magnitude of the effects generally is small to moderate. Nonetheless, these results suggest additional work is needed to identify reasons for attrition and its effects on analytical modeling of crime victimization.

Table 4.8 Logistic regression – response variable: Interview 3 completed (1=yes), independent variables from interview 2*

Variable	$\hat{\beta}$	*Std err*	*p-value*	*exp*$(\hat{\beta})$
Intercept	−0.541	0.032	<.0001	0.623
Age	0.012	0.000	<.0001	1.008
Male	−0.176	0.015	<.0001	0.936
Black	0.062	0.024	0.0099	1.176
Married	0.194	0.016	<.0001	1.256
HHincome	0.054	0.002	<.0001	1.067
eveningsIn	0.043	0.007	<.0001	1.053
Tenure	0.042	0.001	<.0001	1.059
cntrCityMSA	−0.150	0.016	<.0001	0.873
Income MDD	−0.229	0.021	<.0001	0.876
Evenings MDD	−0.323	0.060	<.0001	0.906
Tenure MDD	−0.276	0.092	0.0027	0.713
$n = 107054$				

[*] $1994 \leq$ base year ≤ 1996.

5 Summary and discussion

This chapter describes how to use the National Crime Victimization Survey (NCVS) for individual-level analytical statistical work in a longitudinal format. The final section of the chapter, in particular, illustrates the type of analyses that can be conducted with these working datasets. It replicates part of the work reported in a paper by Janet Lauritsen (2001). Our results closely match Lauritsen's published work. In addition, we find (1) estimated effects of regressors such as gender, race, age and marital status differ noticeably by interview number, (2) sample attrition is significantly related to all the regressors in the Lauritsen model, and (3) the odds of being a victim of a violent crime substantially decline during the period 1973 to 2005, including controls for demographic variables such as age, race and residence in the center-city of an MSA.

Our experience with the Lauritsen model suggests that one promising line of research is to examine simultaneously crime victimization and sample attrition as a linked process. Variations on hazard models designating crime victimization as one (transient) state and sample attrition as another (nontransient) state may be estimated by maximum likelihood (Tuma and Hannan, 1984). As Tuma and Hannan cogently argue, an important advantage of explicitly modeling the hazard (or survival) function is that one can derive implications for related outcome variables that are, in fact, measured. For binary regression (victim versus nonvictim), it is informative to derive the link function from a survival model. This is unlikely to imply a logistic regression. The exponential survival function may be a good starting point due to its simplicity, and it is not consistent with the logistic link function for a binary outcome. A generalization of this approach is to model the number of incidents (e.g., number of victimizations); the exponential survival function generates a poisson distribution for these counts. For purposes of the NCVS, "series crimes" could then be counted as censored and the likelihood equation specified accordingly.

For some purposes, matching household, person, and incident information may be done using the household ID, person ID, and the address interview number. In other instances, it is important to substitute the household interview number for the address interview number. The merge operations needed to produce a sample like that used in the Lauritsen analyses and our replication of it do not depend on correct identification of the household interview number. The household and person ID variables and address interview number are sufficient. However, other analyses, like those reported in Table 4.7 columns 2 through 8, do depend on correct calculation of the household interview number.

The NCVS provides researchers with the opportunity for answering numerous research questions that cannot be examined using cross-sectional data including issues related to repeat victimization, which are paramount to both theory building and crime control policies. When using these data, it is important to recognize that the sampled units in the NCVS consist of residential addresses, not households or individuals. When a household moves out of a sample address and a new household moves in, the new household replaces the old household. An important conceptual distinction illuminated in the chapter differentiates between the (scheduled) interview number of the address and the (scheduled) interview number of the household. These coincide for nonreplacement households, but the household interview number is less than the address interview number for replacement households. These and other features of the rotating panel design of the NCVS must be taken into account, and increase the complexity of preparing the data, for longitudinal analysis.

Glossary

ADier_no Address interview number (variable name)

base year Year when interview occurred (base year)

"by" variable Variable used to match cases in two or more files to be merged

HHier_no Household interview number (variable name)

Panel Subdivision of the NCVS sample rotation. Each rotation is divided into six panels. Interviews of respondents in each panel begin at one-month intervals. January or July for panel 1, February or August for panel 2,…

Reference period Time interval six months prior to day 1 of the month the interview occurred. Respondents are asked to report victimizations that occurred during the reference period

Rotation Subdivision of each NCVS sample. Each sample is divided into six rotations. Interviewing of successive rotations begins at six-month intervals

Sample Addresses identified for inclusion in the NCVS. A new sample is selected every three years. Samples are designated by integer numbers prefixed with the letter J

Scrambled Control Number (SCN) Numeric address identification code, defined within sample numbers

TIS Time in sample, a synonym for ADier_no

References

BJS (2006). Serious violent crime levels declined since 1993. Online summary report. (http://www.ojp.gov/bjs/glance/cv2.htm), accessed April 17, 2006.

Bowers, K.J. and Johnson, S.D. (2005). Domestic burglary repeats and space-time clusters. *European Journal of Criminology*, 2: 67–92.

Conaway, M.R. and Lohr, S.L. (1994). A longitudinal analysis of factors associated with reporting violent crimes to the police. *Journal of Quantitative Criminology*, 10: 23–39.

Farrell, G., Sousa, W.H., and Weisel, D.L. (2002). The time window effect in the measurement of repeat victimization: A methodology for its measurement and an empirical study. In N. Tilley (ed.), Analysis for Crime Prevention. Vol. 14, Crime Prevention Studies. Monsey, NY: Criminal Justice Press.

Farrell, G., Tseloni, A., and Pease, K. (2005). Repeat victimization in the ICVS and the NCVS. *Crime Prevention and Community Safety*, 7: 7–18.

Gabor, T. and Mata, F. (2004). Victimization and repeat victimization over the life span: A predictive study and implications for policy. *International Review of Victimology*, 10: 193–221.

Gallagher, C.A. (2005). Injury recurrence among untreated and medically treated victims of violence in the USA. *Social Science & Medicine*, 60: 627–635.

Lauritsen, Janet L. (2001). The social ecology of violent victimization: Individual and contextual effects in the NCVS. *Journal of Quantitative Criminology*, 17: 3–32.

Lauritsen, J.L. and Davis Quinet, K.F. (1995). Repeat victimization among adolescents and young adults. *Journal of Quantitative Criminology*, 11: 143–166.

Laycock, G. (2001). Hypothesis-based research: The repeat victimization story. *Criminal Justice*, 1: 59–82.

Menard, S. (2000). The 'normality' of repeat victimization from adolescence through early adulthood. *Justice Quarterly*, 17: 543–574.

Outlaw, M.S., Ruback, R.B., and Britt, C. (2002). Repeat and multiple victimizations: the role of individual and contextual factors. *Violence and Victims*, 17: 187–204.

Stevens, T.N., Ruggiero, K.J., Kilpatrick, D.G., Resnick, H.S., and Saunders, B.E. (2005). Variables differentiating singly and multiply victimized youth: results from the National Survey of Adolescents and implications for secondary prevention. *Child Maltreatment*, 10: 211–223.

Tuma, N.B. and Hannan, M.T. (1984). *Social Dynamics*. Orlando, FL: Academic press.

US Dept. of Justice, Bureau of Justice Statistics. NATIONAL CRIME VICTIMIZATION SURVEY, 1992–2004 [Computer file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 2006.

Xie, M., Pgarsky, G., Lynch, J.P., and McDowall, D. (2006). Prior police contact and subsequent victim reporting: results from the NCVS. *Justice Quarterly*, 23(4): 481–501.

Ybarra, L.M.R. and Lohr, S.L. (2002). Estimates of repeat victimization using the National Crime Victimization Survey, *Journal of Quantitative Criminology*, 18: 1–21.

This page intentionally left blank

Chapter 5

The Millennium Cohort Study and mature national birth cohorts in Britain

Heather E. Joshi

The national birth cohort studies, pioneered in Britain, have been rated of enormous importance for both scientific and policy understandings of human behavior. The Millennium Cohort Study (MCS) is the latest of this series of prospective studies of the life course, from 1946, 1958 and 1970, and builds upon them. It adds to the portfolio of longitudinal data available for secondary analysis on the United Kingdom, and adds to the possibilities of cross-cohort and cross-national comparisons.

1 Introduction

Large-scale longitudinal surveys follow individuals through time to chart their experience of political, social, demographic and economic change. Besides such social monitoring, they can also investigate hypotheses about the long-term causes and consequences of experiences, such as disease or educational attainment. The expense of collecting and maintaining these databases often means that they are used to serve many purposes and are utilized for secondary analysis beyond the ideas or imagination of the originators. The prospects of multiple uses, and the difficulty of knowing what topics will interest researchers 25 or 50 years down the line, argue for a broad coverage; on the other hand, to study well specified hypotheses in depth, points towards selectivity, given the twin constraints on the cost of data collection and respondent burden. The Millennium Cohort Study is designed from the outset (unlike its forerunners) to be a multipurpose, longitudinal research resource, with scope for analysis that should interest readers of many disciplines.

Section 2 outlines some features of the three national birth cohort studies started in Britain before 2000 and still ongoing, which formed a template for the study of the Child of the Century (as MCS is known in the field). Other major longitudinal data resources such as the ONS Longitudinal Study, the British Household Panel Study and the Avon Longitudinal Study of Parents and Children are left outside its scope. Section 2 introduces the 1946 birth cohort, followed by the more closely entwined histories of the 1958 and 1970 cohort studies, which are more widely available for analysis. The UK Millennium Cohort Study (MCS) is described in greater detail in Section 3, covering its establishment, design and content from 2000 to 2008. Section 4 discusses the analysis

Table 5.1 British national birth cohort studies: origins and access

<table>
<tr><th></th><th><i>MRC National Survey of Health and Development (NSHD)</i></th><th><i>National Child Development Study (NCDS)</i></th><th><i>British Birth Cohort Study (BCS70)</i></th><th><i>Millennium Cohort Study (MCS)</i></th></tr>
<tr><td><b>Birth year</b></td><td><b>1946</b></td><td><b>1958</b></td><td><b>1970</b></td><td><b>2000–1</b></td></tr>
<tr><td>Initial principal investigators</td><td>James Douglas, 1946–79</td><td>Neville Butler, 1958–69
Mia Kellmer Pringle, 1965–79</td><td>Geoffrey and Roma Chamberlain, 1970–72
Neville Butler, 1973–91</td><td>Heather Joshi, 2000–</td></tr>
<tr><td rowspan="2">Successors</td><td rowspan="2">John Colley, 1979–84
Michael Wadsworth, 1984–2006</td><td>Ronald Davie, 1979–85 John Fox, 1985–89
John Bynner, 1989–2004</td><td>John Bynner, 1991–2004</td><td rowspan="2"></td></tr>
<tr><td colspan="2">Jane Elliott, 2004–</td></tr>
<tr><td>Initial purpose</td><td>Survey of maternity services</td><td>Perinatal mortality survey</td><td>Perinatal conditions with follow-up intended</td><td>Multipurpose, multidisciplinary, longitudinal study</td></tr>
<tr><td>Initial sponsor</td><td>Population Investigation Committee and Royal Commission on Population, Nuffield Foundation, National Birthday Trust Fund</td><td>National Birthday Trust, Royal College of Obstetricians and Gynaecologists</td><td>National Birthday Trust, Royal College of Obstetricians and Gynaecologists</td><td>Economic and Social Research Council, UK government departments</td></tr>
<tr><td>Current funding</td><td>Medical Research Council for core funding, grants from other sources</td><td>ESRC (+MRC*), US NICHD for second generation in 1991, etc.</td><td>ESRC (+NRDC and ESF in 2004)</td><td>As above</td></tr>
<tr><td>Current purpose</td><td>Health over the life course, ageing and its precursors</td><td>Multipurpose studies of the life course, assessment of biomedical outcomes and risk factors</td><td>Multipurpose studies of the life course</td><td>Multipurpose studies of the life course</td></tr>
<tr><td>Availability of data to other researchers</td><td>May be analyzed by collaborators of the study team. Access under review</td><td colspan="3">Anonymized datasets available from the UK Data Archive Biomedical, geographical and other sensitive or disclosive data on special conditions</td></tr>
</table>

* 1958 Cohort Biomedical Study, 2002–7 Christine Power, David Strachan, Bynner/Joshi, Gillian Prior, extended on Wellcome Trust funding to establish a genetic data resource for medical research.

Table 5.2 British national birth cohort studies: research design

	MRC National Survey of Health and Development (NSHD)	*National Child Development Study (NCDS)*	*British Birth Cohort Study (BCS70)*	*Millennium Cohort Study (MCS)*
Criterion	Born in a week of March 1946	Born in a week of March 1958	Born in a week of April 1970	Born over a year* and living in a sampled ward at 9 months
Geographical coverage	England, Wales and Scotland (Great Britain)	Great Britain	UK, but only GB followed up	UK (includes Northern Ireland)
Size of initial sample	5362 Legitimate singletons selected from data on 13,687 out of 16,695 births in the week	17,414 Data collected from 17,733 births in the week. Immigrants with sample birth dates added up to age 16	16,571 17,052 live births in GB. Immigrants with sample birth dates added up to age 16	18,818 Selected from 27,201 children with relevant Child Benefit address and birth months
Weighting	Follow-up of one-in-four children of wives of manual workers and all births to wives of non-manual and agricultural workers	None	None	Over-representation of wards with high child poverty, in Celtic countries, and, in England, with minority ethnic concentration
Frequency of follow-up	20 core follow-ups in 53 years: Approximately every two years until age 26, then 31, 36, 43, 53, 60	7 core follow-ups in 47 years: 7, 11, 16, 23, 33, 42, 46	6 core follow-ups in 34 years: 5, 10, 16, 26, 30, 34	Every two years in childhood 3, 5, 7 planned
		Now planned for 4-year intervals, alternating face to face and telephone, unless further funds become available.		

* MCS sample birthdates between September 2000 and August 2001 in England and Wales and 24/11/00–11/01/02 in Scotland and Northern Ireland

Table 5.3 British national birth cohort studies: data collection and cohort maintenance

	MRC National Survey of Health and Development (NSHD)	*National Child Development Study (NCDS)*	*British Birth Cohort Study (BCS70)*	*Millennium Cohort Study (MCS)*
Sample size at 36, 33 or 30	3322 (36, 1982)	11,407 (33, 1991)	11,261 (30, 2000)	NA
% of eligible	75.5	71.6	71.5	72 (initial sample)
Date of latest contact	1999	2004–5	2004–5	2003–5
Achieved sample at latest contact	3035	9530	9665	15,600
Informants	Mothers, cohort child/member, teachers, school medical officers, school nurses, youth employment officers. Second generation: cohort members or partners when first born aged 4 years and 8 years All women in cohort age 47–54	Mothers, cohort child/member, teachers School medical officer Partners at 33	Mothers, cohort child/member, teachers School medical officer	Mothers, fathers cohort child, teachers, health visitors
		Second generation study on subsample at 33/34 : data from cohort's children and their mothers		
Mode	Interviews, initially by health visitor, postal questionnaires, nurse interviewer visits CAPI in 1999 Clinic data collection feasibility study currently taking place	Interviews, initially by health visitor, paper self-completions, nurse visit at 42 (CAPI/CASI since 2000) Alternate contacts by telephone starting 2004	Interviews, initially by health visitor, paper self-completions, postal questionnaire at 26, CAPI/CASI 2000 Alternate contacts by telephone starting 2008	CAPI and CASI interviews from outset

Table 5.3 (Continued)

	MRC National Survey of Health and Development (NSHD)	*National Child Development Study (NCDS)*	*British Birth Cohort Study (BCS70)*	*Millennium Cohort Study (MCS)*
Tracing	Team of tracers consult various registers, e.g. electoral roll, phone directory, vehicle and driving licence authority, National Health Service Central register will forward enquiries to GP Interviewers attempt to trace movers in field. Stable addresses collected at interview, and followed up on phone by tracers if needed. Cohort members invited to keep in touch. Week of birth a key identifier.			As other studies but also includes updates on addresses from Department for Work and Pensions
Feedback	Annual birthday card with feedback information and invitation to update address			Feedback documents sent in batches around times of birthdays
	Cohort member website			

potential of the latest study. To complete this introduction, Tables 5.1, 5.2 and 5.3 present a cross-cutting synopsis of the history, data accessibility, design and data collection methods of the four national birth cohort studies.

2 The heritage of birth cohort studies in Britain

2.1 The MRC National Survey of Health and Development: 1946 Cohort

As shown in Table 5.1, the first national survey of births took place in 1946 addressing the state of maternity services on the eve of the introduction of the National Health Service (Joint Committee, 1948). Because of the urgency of the situation and cost constraints, the survey took all the births in one week (in March) and data collected by health visitors in the home a few weeks after delivery. The maternity study covered all the births in Great Britain where the authorities cooperated (13,687 out of the possible total of 16,695, Table 5.2). It made recommendations about maternity care published in the Joint Committee's report (1948). The Director, Dr James Douglas, initiated a follow-up in 1948, of a weighted subsample of 5632 cases from the original 13,687. Unforeseen at the time, this continues into the 21st century. There have been 20 follow-ups across childhood and adulthood up to 2000 (see Tables 5.2 and 5.3 and Wadsworth et al., 2003). A further sweep is being prepared for age 60 in 2006. There were approximately 3600 cohort members remaining in the study at the 1999 interview. The 1946 study has pioneered methods of keeping in touch with cohort members, including sending them a birthday card.

The study is known as the MRC National Survey of Health and Development to reflect its major funding from the Medical Research Council, which since 1962 has permitted the continuation of the study over the cohort's adult years, directed since 1984 by Professor Michael Wadsworth at University College, London. This funding of data collection and analysis together follows a model customary for medical research in Britain. The dataset is not conceived as a general data resource and there are features of the consent given by informants which may limit its uses to specified purposes. Researchers outside the MRC Unit at University College, London, have been able to work on the study via collaborative arrangements, and wider data access is under review.

Classic findings from the childhood years are reported in books by Douglas (1964) and Douglas et al. (1968). Further findings have been synthesized by Wadsworth (1991) and, for example, presented in collections of papers on life course epidemiology and women's health (Kuh and Ben Schlomo (eds), 2004, and Kuh and Hardy (eds), 2002). Further information about the study and publications arising from it can be found on its website (www.nshd.mrc.ac.uk). The volume edited by Ferri, Bynner and Wadsworth (2003) brings together findings on the 1946 cohort up to age 43, with similar results up to age 42 and 30 from the 1958 and 1970 birth cohorts, respectively. It also summarizes the methodology of all three studies.

2.2 The National Child Development Study: 1958 Birth Cohort

The 1958 cohort study also features in Tables 5.1, 5.2 and 5.3. The birth study was again about perinatal conditions (Ferri, 1998). Like the 1946 exercise it took all the births in a week of March as the target and health visitors collected the data. There was also no initial plan to establish a longitudinal study (despite the example of eight follow-ups of the 1946 children by then). That opportunity came in 1965 when the cohort sample was revived to provide evidence for the Plowden Enquiry on primary education, which reported in 1967. This time there was no subsampling, indeed the cohort was augmented. In 1965, 15,425 seven-year-olds were found, of whom 15,051 had been in the initial sample of 17,415 and 374 were subsequent immigrants, born in the survey week, identified by their birth dates in school records (Plewis et al., 2004).

The survey became known as the National Child Development Study (NCDS) and was based under the joint direction of its founder, Neville Butler (by then Professor of Child Health at Bristol University), and Mia Kellmer Pringle, at the National Children's Bureau from 1965. Under the Bureau's auspices, and with various ad hoc sources of funding, there were three childhood follow-ups: NCDS1 at age 7, NCCDS2 at age 11, and NCDS3 at age 16 (Davie et al., 1972, Fogelman, 1983). Each time immigrants were recruited. Information was collected from mothers, teachers, school medical services and, at age 16, from the cohort member as well. Examination results were added. The data were made available to the research community via the ESRC Data Archive from 1983. The National Children's Bureau oversaw the collection of the first survey of the cohort as adults, NCDS4 at age 23. This round of data collection was largely funded by government departments but there was no commitment to longer term continuation.

The size of the sample from which some information was collected was around 14,000 at age 16 and somewhat over 12,000 at age 23, representing a longitudinal response rate out of the original cohort of 87% and 76% respectively (Plewis et al., 2004). The proportions of the original cohort with complete information are somewhat lower, depending on which variables are of interest.

In 1985 the National Child Development Study (as it is still known), transferred to the Social Statistics Research Unit at City University, directed by Professor John Fox, who was succeeded in 1989 by John Bynner. In 1998, Professor Bynner took the unit, and the two cohort studies for which it was by then responsible, to the Institute of Education, in the University of London. The team still operates there, renamed as the Centre for Longitudinal Studies (CLS).

Back in the 1980s John Fox faced the challenge of raising funds for a fifth sweep

(NCDS5). The Economic and Social Research Council (ESRC) joined government departments to finance a survey in 1991 of cohort members at age 33 and their partners. Its design, as with future CLS surveys, was developed in consultation with the scientific community. It also includes a survey of the children of one-third of the cohort (and their mothers). This second generation study was financed by the US National Institute for Child Health and Development (NICHD) and contained developmental measures comparable with the second generation module of the National Longitudinal Survey of Youth (NLSY) in the US. A public use data set of 11,407 cases with at least some information (10,986 of whom were also in the birth survey) was made available through the Data Archive, and its preliminary findings outlined in a 'sourcebook' (Ferri (ed.), 1993).

2.3 The British Birth Cohort Study of 1970(BCS70)

During the 1990s the development of the 1958 cohort became increasingly integrated with that of the 1970 cohort study, so this is a good point to bring in the next study featured in Tables 5.1 to 5.3, and rewind history to 1970, when the third nationwide birth cohort study was launched of all births in a week of April 1970, under similar auspices as in 1958 (Ferri, 1998). Health visitors, again, collected data on 16,571 births (out of a possible 17,287 in Great Britain, Plewis et al., 2004). The birth survey also included, for the first time, some 600 births in Northern Ireland, but they were not followed up in the deteriorating political situation. This time there was always an intention to follow up and subsamples were followed up at 22 and 42 months. The Child Health and Education Study (CHES) was established to pursue a broad front of social and educational issues, as well as medical, in the Department of Child Health, University of Bristol, under the direction of Neville Butler. There was a full survey at age 5 (Osborn et al., 1984) and another CHES survey at 10. The age 16 follow-up was conducted by a charity, the International Centre for Child Studies (ICCS), with funding from a large number of sources, thanks to Neville Butler's dual talents as a scientist and fundraiser. Known as Youthscan, this survey collected a great variety of data from the young people, their parents and schools, some of which remained inaccessible for a number of years due to a lack of resources for data management.

In 1991 the unit at City University took responsibility for the BCS70 (as it became known) and deposit of anonymised data for public use in the ESRC archive (Ekinsmyth et al., 1992). A one-in-ten sample was surveyed at age 21 on literacy and numeracy (and complemented by a similar 10% survey of the 1958 cohort at age 37). In 1996 the opportunity of ESRC funding arose at short notice to approach the whole 1970 cohort, at age 26, with a postal questionnaire about their lives since leaving school. Given the short time available to trace addresses and follow up nonresponders, the 9006 returns were not only interesting in their own right (Bynner et al., 1997) but encouraged the view that this cohort still had potential as a longitudinal data resource.

2.4 NCDS and BCS70 under shared direction

Meanwhile several reviews of longitudinal data endorsed the unique strengths of the accumulating resource and recognized that uncertainties surrounding future funding had not helped rational planning of good data quality. In 1998 the ESRC established a national strategy for longitudinal data. It was agreed that the 1958 and 1970 cohorts should follow a "forward plan" of a sweep every four years, starting with a very similar interview survey of both of them in 2000. Thereafter, alternative data collection, starting with NCDS in 2004, would be through the cheaper telephone mode, unless additional funding is raised. In 2000, the Centre for Longitudinal Studies won the contract for the first survey of the Millennium Cohort study (see fourth column of Tables 5.1, 5.2 and 5.3 and below). In 2004, the ESRC consolidated its funding of the collection and development of the three cohort studies of 1958, 1970 and 2000, up to 2010, making CLS an ESRC Resource Centre. A governance structure was set up to combine the advice of scientific experts with a steering group representing the major funders and reporting to higher level bodies, including the ESRC committee for National Strategy on Longitudinal Data.

The "2000" interview surveys of both the 1970 and 1958 cohort (which actually started in December 1999) used a more or less common interview schedule designed also to permit comparisons of both cohorts between the ages of 30 and 42 respectively, and with the data gathered on the 1958 cohort at 33 (Bynner et al., 2000). The achieved sample sizes, each well over 11,000, meant that BCS70 had recovered cases since the postal sweep and that there had been relatively little net attrition in NCDS. An overview of the results, also making comparisons with the 1946 cohort, can be found in Ferri et al. (ed.) (2003), as mentioned above.

The next major data collection from the 1958 cohort was in 2003–4 around age 45, when biomedical data and specimens were collected by research nurses. Measurements include blood pressure, lung function, blood and saliva samples, hearing and vision tests, height, weight and psychological indicators. This study was initiated by Professors Christine Power and David Strachan, and funded by the Medical Research Council, to assess biomedical outcomes and risk factors. The project has a team of specialist collaborators, medical scientists who, on the MRC model, are funded to analyze the results in the first instance. There is intended to be some form of wider access to the data for other analysts, probably after 2007. The biomedical project has also established a collection of genetic evidence derived from immortalized cell lines, which have been generated under funding from the Wellcome Trust. Access to this material, for purposes of medical research, is controlled by the MRC WT Oversight Committee.

In 2004–5 there was another interview survey of the 1970 cohort at age 34, enhanced by a survey of the children of one cohort member in two. The mother and child survey was financed by funds raised by the National Research and Development Centre on Adult Literacy and Numeracy (NRDC), largely from the European Social Fund. NRDC also supplemented the resources available from ESRC for the survey of BCS70 adults at 34 to permit a module on dyslexia. The second generation study administered cognitive assessments to children old enough to attempt them, like the NCDS second generation study at 33. The number of children surveyed, just over 5000, reached expectations but overall response from adults in the age 34 survey fell below 10,000 (74% of the restricted target issue for fieldwork) reflecting difficulties in finding movers' addresses.

There was, at about the same time in 2005, a telephone survey of the 1958 cohort lasting about 30 minutes and therefore collecting less than their previous 90-minute interview, which showed a similar level of response, n = 9534, to the survey of BCS70 that year (9665, see Table 5.3), but a higher proportional success rate (81%) in reaching addresses that were actually issued for fieldwork.

In 2004, after John Bynner's retirement, Dr Jane Elliott took over as the Research Director of the NCDS and BCS70. In 2006 she was presiding over the preparation of material from the 2004–5 studies for deposit in the UK Data Archive and preparing for the next round of surveys on these adult cohorts in 2008. Given the growing mass of information collected on these the cohort members' lifetimes (well over 13,000 variables) consideration is being given to improving the disclosure control associated with data access.

3 The ESRC Millennium Cohort Study

3.1 Gestation and metamorphosis of the first survey

The history of the Millennium Cohort Study starts with the nonoccurrence in 1982 of any fourth in a twelve-year sequence of national birth cohort studies. The political and economic climate then did not favor large investments in social data collection. It was to be another 16 years before the ESRC adopted its longitudinal data strategy. Then the government decided it would include a new cohort study among the activities to celebrate the turn of the Millennium. The ESRC, which was to be the main funder, drew up a specification for the study. It was to have more social and economic content and less of a medical focus than the initial surveys of its predecessors. Data were to be collected by computer-assisted means by professional interviewers, rather than by imposing on the goodwill of health visitors, unlikely still to be possible. The relative scarcity of trained interviewers told against attempting to recruit all the births of one week again. Another consideration in favor of spreading births throughout the year was that it would reveal any variation by season of birth, excluded by the earlier design. It was stipulated that at least some of the cohort should be born in 2000 and that the first survey should be within the first year of life, as close to age 6 months as possible. The invitation to tender for the principal investigator (PI) role was not published until late February 2000, with unusually short notice to submit. In May 2000, it was announced that the Centre for Longitudinal Studies had won the PI contract, with the author as MCS Director. But already nearly half-way through the Millennial year, it was still necessary to specify the tender for fieldwork and consider bids for that before work could start in earnest. The National Centre for Social Research (NatCen) was appointed as the fieldwork contractor in October 2000.

By that time there was little choice but to start fieldwork in 2001. The cohort births were fixed in the 12 months from September 1, 2000, and the interview age at 9 months, i.e., to start fieldwork in June 2001. Even so, the survey development phase, with two pilot surveys, proceeded at a breakneck speed, maintaining the tyrannical pace from which the study never seems to escape.

Two major features of the survey design were established during the earlier months of the PI's work. The first was the sampling scheme devised by Ian Plewis, which simultaneously permitted the over-representation of certain groups of particular scientific or policy interest, and provided a structure making it possible to analyze neighborhood and community on child development and family wellbeing (Plewis, 2004). The subpopulations of interest were children in poor families, children from minority ethnic groups, and, especially given supplementary funding from the devolved administrations of these countries, the inhabitants of Scotland, Wales, and Northern Ireland. The geographical unit in which these populations were identified is the electoral ward, an administrative entity relevant to elections (but not any other particular service provision) averaging around 5500 inhabitants. There are 11,000 of these in the UK as a whole. There were data available on families receiving low-income benefits by ward in 1998, which were used to split each county into two strata, disadvantaged and nondisadvantaged (sometime rather misleadingly labeled “Advantaged”since this comprises all but the most disadvantaged). The cut-off point for the child poverty index was that the ward had more that 38.4% of the children in a 1998 database living in “poor” families. This cut-off (bottom quartile of wards in England and Wales) accounted for nearly half the wards in the smaller countries of Wales, Scotland, and Northern Ireland, but less than one quarter in England. In England, where most of the minority ethnic groups live, there is a third stratum, actually identified before the other two, of wards expected to contain high concentrations of ethnic minority population. The criterion had to be based on evidence from the 1991 census, which indicated places redrawn to 1998 boundaries, with more than 30% of the population of Black or Asian ethnic identity in 1991 (where the median for these places was 51%). With two strata in each of the Celtic countries and three in England, there are nine strata, within which wards are selected with different known probabilities, offering the greatest over-representation to the ethnic stratum in England, followed by the disadvantaged stratum in Wales. 398 wards were selected and within them all births on specified birth dates (see below "Qualifying Dates of Birth") were eligible for the survey, provided the child was resident in the ward at age 9 months.

The second decision involved sampling frames and the way in which respondents could be approached. The first sampling frame was the list of wards on boundaries that could be matched to child poverty and ethnic data, but at the second should locate the children in selected wards who had the selected birthdays. One possible source of names and addresses and dates of birth was the registration of births. This avenue was abandoned because the recruitment procedures involved parents providing a written opt-in consent before their names could be released to the survey. There would not have been time to follow-up initial nonresponders and there was concern that an opt-in would bias the sample away from people with poor language or literacy skills.

An alternative sampling frame was offered by the then Department of Social Security, which was joining the consortium of government departments offering major supplementation to the funding of the survey. This permitted access to the Child Benefit Register. Child Benefit is almost universally claimed and its records contain both a date of birth and an address likely to be reasonably up to date. Furthermore, the Department operated an opt-out rather than an opt-in approach when using the records to recruit research samples. The Department sent a letter to parents with an eligible address saying that the Department would pass their name and address on to the survey, unless the parent provided written or telephoned notice that they would prefer to opt out. In the event, 7% opted out, a further 3% of cases were held back as potentially sensitive or already issued to another survey, which combined with 13% refusals in the field compares favorably with the 30% or so pre-fieldwork loss feared by the opt-in route. The opt-out route raised questions when the plan was presented to the NHS Research Ethics Committee, who felt it might be coercive. It was agreed that interviewers should establish that the families really were giving informed consent at an introductory visit before proceeding.

The confirmation of government funding to supplement the resources originally committed by the ESRC to the Millennium Cohort during its design phase increased the scope of the study in several ways. One was to increase the target achieved sample from 15,000 originally specified to over 20,000. Funding from the National Assembly of Wales doubled the target number of cases in that country to 3000. There were also boosts of 1000 and 500 in Scotland and Northern Ireland respectively, funded by the devolved governments of those countries. The sample of disadvantaged wards in England was boosted to provide more control cases for the National Evaluation of Sure Start in England, an integrated program of services for young children. Government funding also paid for an extension of the length of contact time with families by 15 minutes in the first survey, to 105 minutes overall. The government funding permitted several enhancements to the interview data, notably the linkage of external data (from hospital episode statistics, birth registration) for those (the vast majority) who gave permission and could be matched. Other enhancements include: a postal substudy of mothers whose cohort child was born through assisted fertility; another postal survey of health visitors about the services available to families with young children in the study wards; and the assembly of other ecological indicators about those places. Crucially, and unlike the ESRC award, the government funding provided funds for in-house reporting and analyzing the study up until 2005, during which period it also earmarked funds to contribute to the collection of data for the second survey.

3.2 Qualifying dates of birth

Before the story of the Millennium Cohort can proceed it is necessary to set out the dates of birth, which qualify a child for membership of the cohort. For families resident in England and Wales, they run from September 1, 2000 to August 31, 2001, which fulfills the requirement of being spread over a year, coincidentally the dates normally separating the age cohorts passing through the school system. The survey could not start with September births in Scotland and Northern Ireland, because of an existing government survey drawing respondents from those families. It was decided to avoid respondent overload by postponing the start of MCS fieldwork. The cohort started in these countries with births from November 24, 2000, the start of the fourth batch of 4-week birth cohorts who formed the waves in which the sample was issued (Shepherd et al., 2004). The closing date for the birth dates for the cohort in these countries should have been November 23, 2001, but the sample of birth dates was extended, by an extra six weeks to January 11, 2002. This decision was taken in mid-fieldwork to compensate for the shortfall of cases being issued. This resulted from the actual number of births falling below the number expected when the sample was designed. In fact, 2001 turned out to be an all-time low for British fertility. By the time the "birth dearth" became evident it was not possible to boost the sample by selecting more wards; the only option was to select more birth dates. The sponsors in England and Wales were content to live with the shortfall in numbers, but the samples in Scotland and Northern Ireland were anyway smaller, so the option of extending birth dates was taken up. The rules for school entry and dates of school years in Scotland and Northern Ireland are not the same as in England and Wales, so the feature that members of the cohort do not belong to a single school year is magnified, but not created, by these variations in the cohort birth dates.

3.3 Data collection over time

With its confirmation in late 2001 of funding for the second survey at age 3, ESRC ensured that the study would, as intended, become longitudinal. The institution of CLS as a Resource Centre in 2004, mentioned above, provided funds for the third and fourth surveys at age 5 and age 7. The fieldwork for these two surveys mostly occurs in 2006 and 2008 respectively. Any further follow-up, which might preserve the two-year rhythm, with a survey at age 9, or mimic the 10-year-old survey of BCS70 (in 2011) or the 11-year-old survey of NCDS (2012), has yet to be decided, but there is, in principle, the intention to follow the cohort further into adulthood.

The fact that the sample birth dates extend over more than 16 months has implications for the duration of fieldwork in each survey, subsequent data release, and the frequency with which they can be repeated. In the first survey, interviews mostly took place as intended, very close to the time the child was aged 9 months, but this still meant that fieldwork spread from June 2001 to January 2003. By that time preparations for Sweep 2 were well in hand. Designed to reach families when the child was between 36 and 39 months of age, fieldwork should have run from September 2003 to the end of February

2005, but the finish slipped to the beginning of May 2005. Fieldwork periods in the next two sweeps, age 5 and age 7, will be less spread out in England and Wales, by dint of interviewing at a greater spread of child ages, but in Scotland and Northern Ireland the split of the sample over two school years will extend the collection and complicate the release of data. A preliminary version of the first survey (MCS1) was released to the UK Data Archive in mid 2003. Another version was released in the spring of 2006 with comparable data from Sweep 2, with variable names reissued to facilitate longitudinal analysis.

3.4 Content

The broad topics covered by the first four surveys of MCS are summarized in Table 5.4. MCS1 conducted interviews with each resident parent, taking many questions from those recently put to members of the 1958 and 1970 cohorts. The main informant was almost invariably the natural mother, supplemented by the main informant's partner, usually the natural father if he was present and willing. Each parent also had a self-completion instrument containing some more sensitive material. The whole encounter took 105 minutes on average, about 75 minutes with the mother and 30 with the father. She was asked about a number of topics including pregnancy and delivery, which were not repeated for the father. This is a long time for an interviewer visit (or visits), taking even longer when questions had to be translated, but on the whole parents and interviewers seemed to enjoy talking about the baby.

The age 3 survey repeated the parental interviews with similar but not identical content, over a somewhat shorter total span of time, because this was the first survey of the cohort to take direct measurements from the cohort children themselves. These consisted of anthropometry and cognitive assessments (details in

Table 5.4 UK Millennium Cohort Study: content of first four surveys

Respondent	*Mode*	*Modules and subtopics*	*Surveys, by age of cohort*	
Main/Partner	Interview	Household (inc. ethnicity and language)	9m,3,5,7	
			Mother/main	Father/Partner
Parents	Interview	Non-resident parents	9m,3,5,7	
		Pregnancy, labour and delivery	9m	
		Father's involvement with child		9m,3,5,7
		Child's health and development	9m,3,5,7	
		Childcare	9m,3,5,7	
		Early education	3,5	
		School	5,7	5,7
		Grandparents and friends	9m,3,5,7	9m,3,5,7
		Parent's health	9m,3,5,7	9m,3,5,7
		Employment	9m,3,5,7	9m,3,5,7
		Family income	9m,3,5,7	
		Parental education/skills	9m,3,5,7	9m,3,5,7
		Housing and local area	9m,3,5,7	
		Interests and time with child	9m,3,5,7	9m,3,5,7
		Older siblings	3,5	

Table 5.4 (Continued)

Respondent	*Mode*	*Modules and subtopics*	*Surveys, by age of cohort*	
	Self-completion	Child's temperament and behavior (inc. SDQ from 3)	9m,3,5,7	
		Relationship with partner	9m,3,5,7	9m,3,5,7
		Previous relationships	9m,3,5,7	9m,3,5,7
		Domestic tasks	9m,3,5,7	
		Parenting	9m,3,5,7	9m,3,5,7
		Previous pregnancies	9m	
		Children living elsewhere		9m,3,5,7
		Mental health	9m,3,5,7	9m,3,5,7
		Drug use, domestic violence	3,5,7	3,5,7
		Alcohol problems	9m,3,7?	9m,3,7?
		Attitudes	9m,3,5,7	9m,3,5,7
Child	Cognitive assessments	Bracken Basic Concept Scale	3	
		BAS naming vocabulary		
		BAS picture similarities	3,5…	
		BAS pattern construction	5…	
		Sally and Anne	5	
	Anthropometry	Height and weight	3,5….	
		Waist circumference	5…	
	Biological sample	Immunology of oral fluid	3	
Older sibling	Self-completion	(England only)	3,5	
Interviewer	Observations	Home environment	3,5 …	
		Neighborhood	3	

Notes
9m = 9months. Other ages in years
The modules are not all repeated in their entirety across sweeps.
Some of the original questions are only put in subsequent sweeps to new informants.
The exact composition of child assessment at Sweep 4 is not yet decided.
BAS is British Ability Scales, 'Sally and Anne' is a Theory of Mind test.

Table 5.4), which the field force was specially briefed to administer. Similar data collection has also been successfully carried out by survey interviewers in the second generation studies of NCDS and BCS70. Another innovation at age 3 was collection of data about older siblings, and in England from some of these children themselves, if aged 10–15, in a self-completion questionnaire. Broadly similar assessments continue in the age 5 survey and there are plans for age 7, with age-appropriate changes. Questions about older siblings were repeated, probably for the last time at 5, but not the self-completion. From age 5, anthropometry includes the measurement of the waist as well as height and weight, in the belief that 5-year-olds might be more cooperative than 3-year-olds at yielding this key information for the study of obesity. At age 3 the children also provided a sample of oral fluid, for an investigation of immunities to test the hygiene hypothesis about allergy and asthma, strictly not for any other purpose. There is no further specimen collection in the surveys at age 5 and age 7, but a biomedical follow-up at a later age is under consideration. The final elements of the age 3 survey (of which vestiges only remain at age 5) were two observations made by the interviewers: conditions inside the home and in the neighborhood outside.

Although some broad themes in the parental questionnaire are shown in Table 5.4 as continuing threads across each survey, the actual questions may not be identical or as many at each sweep. Some information is fed forward to ensure continuity. New informants, mostly new partners, are given extra questions to fill in essential data, which is not otherwise repeated. Details of the instrumentation for datasets in the public arena are available on the CLS website and from the UK Data Archive.

3.5 Response and cohort maintenance

While any survey is judged by its success at achieving a high and unbiased response rate, the issue of achieving and maintaining response is of special concern to the gatherers and analysts of longitudinal data. As indicated in Table 5.3, the mature cohorts had maintained, at least until 2000, response rates into adulthood of over 70% of the eligible population (i.e. excluding those who are known (or thought) to have died or emigrated). For details, readers are referred to Wadsworth et al. (2003) and Plewis et al. (2004). The 1946 response rate is based on the 5632 follow-up sample and does not allow for initial loss when the birth survey was done. For the MCS this item alone brings the response rate at the first survey also to the 72% mark (Plewis, 2004). The success of the fieldwork operation is more often judged by response out of the sample actually issued. For the first MCS survey this was 81% (like NCDS in 2004–5), although the unissued eligible cases for MCS were not found or withheld by the DSS, and the unissued eligible cases for the mature cohorts consist of permanent refusals and those thought to be beyond hope of tracing (given available resources). CLS issues a technical report on sampling with each deposit of data to the archive, estimating and analyzing nonresponse rates, discussing item nonresponse, potential biases, and the possibility of weighting to correct for nonresponse bias. These reports are not yet available for the most recent surveys, where there has clearly been substantial attrition, but also some recuperation of informants who had previously not responded, including a group of "New families", missed by the DWP at the first MCS survey but interviewed in the second.

The technical report on sampling, which accompanied the deposit of the first Millennium Cohort survey (Plewis, 2004) deals with response and potential bias in the drawing of the original sample (noteworthy is the low response from the ethnic minority and Northern Irish wards). The analysis of the 3500 cases lost to the survey between the first and second surveys, is not ready at the time of writing, nor is the analysis of the disappointingly high losses (about 1500 cases per cohort) in the 2004–5 rounds of the 1958 and 1970 cohorts. This has not been for want of efforts to trace (see Table 5.3). It may be that there has been an increase in residential mobility and/or a diminution of the willingness of providers of address information to divulge personal details, given the new provisions of the Data Protection Act protecting privacy. It is not easy to draw a firm line between those unproductive cases who have refused and those with whom no contact has been made. Mobility is not necessarily a problem for follow-up if the cohort member keeps in touch.

Hitherto, the consensus has been that attrition has biased the 1970 and 1958 cohorts somewhat away from less advantaged and less able individuals. This means that the datasets should be treated with caution as sources of cross-sectional estimates of prevalence, but as sources of evidence for models of longitudinal processes they should still be useful provided the analysis controls for variables which are correlated with survey loss, and the possibility of attrition bias is taken seriously on a topic by topic basis.

The secret of successful longitudinal data collection is the goodwill of the cohort members (or their families). The surveys rely on their cooperation not only in giving interviews but in keeping in touch and updating their addresses. The rules of the NHS Research Ethics regime, which must give clearance to the studies, forbids giving material incentives to adults, although it does permit a small present for the cohort child (and a smaller one for their siblings). The only incentive that can be offered the cohort members to continue their cooperation is to let them know that their participation is important, and that their information is being used to good ends. They are offered as much feedback as we can afford, via an annual mailing and a dedicated website, but the budget could only run to offering a discount on the purchase of the book about the "Children of the New Century", and an 8-page magazine-style summary for all the MCS parents.

4 Findings and scope for analysis of the Millennium Cohort

There are over 1000 publications using data from the cohort studies reported to the CLS, listed on the CLS website www.cls.ioe.ac.uk, along with a selection of references to key publications, to which readers are referred as space precludes extensive referencing here. Health-related research on NCDS has recently been reviewed by Power and Elliott (2005). It famously includes establishing the consequences for the child of smoking in pregnancy, linking childhood disease to adult outcomes, discovering which conditions persist and which are outgrown. Work on "health inequalities" has shown the relatively small extent to which socioeconomic variations in outcomes are governed by health selection and the greater extent that risk and disadvantage accumulate across the life course.

The first three cohort studies have also documented the intergenerational transmission of educational advantage, and contributed to a debate on social mobility. They have been used to estimate returns to investment in higher education and the gender premium in pay. They provided key evidence on failure to acquire the basic skills of literacy and numeracy by adulthood. Behavioral as well as cognitive indicators have been shown to play an independent role in predicting outcomes, including crime and adult mental illness. Family disruption in the parental generation predicts faster formation and dissolution of partnerships by the next generation. In due course, the MCS will be able to show whether these patterns persist or change for those growing up in the 21st century, although this will take several decades to materialize.

Initial findings from the first survey of MCS were outlined in Dex and Joshi (2004), which is a set of descriptive tables designed to stimulate further analysis, including external users of the public dataset. A more considered set of essays were collected by the same editors, Dex and Joshi (2005), from the team of collaborators who had helped put the survey together. Other publications using the dataset are beginning to appear and be listed on the CLS website.

The first survey of MCS has already set out the diverse initial circumstances from which the "Children of the New Century" are setting out on life. The range of data already collected and soon to be collected in childhood, provides a richer resource to investigators of the circumstances and outcomes of contemporary children, than did the 1958 and 1970 studies in their early years. The nine months survey permits analysis of the interrelation of the health of the child, health of both parents, social, demographic, economic and attitudinal variables, neighborhood type, and ethnicity. Information on ethnic group is effectively absent from the earlier cohorts, but MCS has nearly 3000 families from minority ethnic groups, tightly concentrated in the areas where it was expected to find them.

Poor families were also concentrated in the areas selected to over-represent them, but there were relatively more outside such areas than ethnic minorities outside the selected wards and more non-poor within poor areas than white respondents in the minority ethnic areas. The minority ethnic families tended themselves also to be economically disadvantaged, but to display cultural diversity across the different groups. Other correlates of economic disadvantage were young motherhood and lone motherhood. Of the (weighted) sample, 16% were mothers without a (regularly) resident partner. The questions revealed a spectrum of degrees of contact between the cohort child and their father, nonresident, semi-resident and present in the household (Kiernan and Smith, 2003). Fathers' involvement in parenting the cohort child began for 86% of them with being present at the birth. This is a marked contrast to the earlier cohorts where such information was not even recorded. On other aspects of childbirth it is possible to make some more precise inter-cohort comparisons: 22% of MCS births were caesarean sections, twice the rate in 1970; only 2% of the births were at home, rather than hospital, compared with 42% in the 1946 cohort and 35% in 1958 (Dex and Joshi, 2005).

The cohort studies have also been used to chart the increase over time of mothers' paid work (Hansen et al., 2006). MCS documents this trend having reached the point where half of all mothers were employed by the time the cohort child was aged 9 months (this took about 5 years after 1970 and 7 years after 1958). As a reflection of this change, and of associated policy interest, it includes far more information on childcare than its predecessors. It also illustrates that there is still a marked variation in mothers' attachment to the labor market, greater for the more educated, older women in two-parent families, and particularly low for certain ethnic groups (Pakistanis and Bangladeshis among the larger minorities), while Black mothers had the highest rates of full-time employment.

Both first and second generation studies of the previous cohorts have been influentially used to demonstrate adverse effects of family poverty on child development. This has helped inspire the policies such as Sure Start and the Children's Fund, which MCS is now being used to monitor. The abolition of child poverty and the support of family life in general is explicit government policy. This contrasts with the very different policy regimes and economic circumstances obtaining when the previous cohorts were children. There is already evidence from the past of relationships between child outcomes and family structure, mother's employment, migration, and neighborhood type (e.g., McCulloch, 2006) which can be explored in greater depth with the richer data collected in MCS.

5 Conclusion

The British national birth cohorts have not only established a tradition in the UK but they have also had their imitators in several other countries around the turn of the Millennium. There are national child cohort studies already underway in Canada, USA, Australia, Denmark, and Norway, with major studies being planned in the USA, France, and Ireland. None are exact replicas of the Millennium Cohort Study, but there are sufficient similarities to permit cross-country as well as cross-cohort comparisons.

Whether the British tradition will be able to sustain existing studies and support any new ones in later decades of the 21st century will depend on maintaining a symbiosis between informants, funders, and users. Potential users of all the CLS cohort studies are invited to register on the Centre's website, www.cls.ioe.ac.uk for up-to-date information, listing of in-house and user publications, detailed documentation, particulars of substudies, and news of data deposits and training events. Users of cohort study data are requested to observe their obligation to report the publications that the studies have yielded for dissemination on this website. The future of the studies depends not only on maintaining a good response rate from cohort members, but also from the scientific community.

Glossary

BCS70 1970 British Cohort Study

CHES Child Health and Educational Study (name of BCS70 at ages 5 and 10)

CLS Centre for Longitudinal Studies, the custodian and curator of the latest 3 national cohort studies, an ESRC Resource Centre since 2004, at the Institute of Education, University of London

DSS Department of Social Security, later Department for Work and Pensions

ESF European Social Fund

ESRC Economic and Social Research Council

Health visitors Community nurses charged specifically with domiciliary care of young families

ICCS International Centre for Child Studies. The charitable organization which sponsored BCS70 (and research on NCDS) from 1983, Director, Neville Butler

MCS Millennium Cohort Study (funded mainly by ESRC, field name Child of the New Century)

MRC Medical Research Council

NCB National Children's Bureau

NCDS National Child Development Study (1958 birth cohort)

NICHD National Institute of Child Health and Development (US)

NRDC National Research and Development Centre for Adult Numeracy and Literacy

NSHD National Survey of Health and Development (MRC survey of the 1946 birth cohort)

ONS Office for National Statistics. Leads consortium of government departments co-funding MCS

SSRU Social Statistics Research Unit, City University, London, the previous home of NCDS and BCS70

Youthscan The name given to the BCS70 survey at 16

References

Bynner, J., Ferri, E. and Shepherd, P. (1997). *Twenty-something in the 90s: Getting on, Getting by; Getting Nowhere*. Aldershot: Dartmouth Press.

Bynner, J., Butler, N.R., Ferri, E., Shepherd, P. and Smith. K. (2000). NCDS6/BCS 2000 The design and conduct of the 1999–2000 surveys of the National Child Development Study and the 1970 British Cohort Study. CLS Cohort Studies Working Paper No.1 www.cls.ioe.ac.uk/library.asp?section=00010001000600060019&page=6J

Davie, R., Butler, N., and Goldstein, H. (1972). *From Birth to Seven*. London: Longman.

Dearden, L., McIntosh, S., Myck, M. and Vignoles, A. (2002). The returns to academic and vocational qualifications in Britain. *Bulletin of Economic Research*, 54(3): 249–274.

Dex, S. and Joshi, H. (eds) (2004). *Millennium Cohort Study First Survey: A User's Guide to Initial Findings*. London: Centre for Longitudinal Studies, Institute of Education. www.ioe.ac.uk/bedfordgroup/publications/userguide.pdf

Dex, S. and Joshi, H. (eds) (2005). *Children of the 21st Century: From Birth to Nine Months*. Bristol: The Policy Press.

Douglas, J. W. B. (1964). *The Home and the School*. London: MacGibbon and Kee.

Douglas, J. W. B., Ross, J. M. and Simpson, H. R. (1968). *All Our Future*. London: Peter Davies.

Ekinsmyth, C., Bynner, J., Montgomery, S. and Shepherd, P. (1992). An integrated approach to the design and analysis of the BCS70 and NCDS inter-cohort analysis. Working Paper 1, CLS, Institute of Education. www.cls.ioe.ac.uk/library.asp?section=00010001000600060012

Ferri, E. (1993) *Life at 33*. London: National Children's Bureau.

Ferri, E. (1998). Forty years on. Paediatric and Perinatal Epidemiology. Vol 12, Supplement 1: 31–44. Special issue in Honour of Professor Neville Butler.

Ferri, E., Bynner, J. and Wadsworth, M. (eds) (2003). *Changing Britain, Changing Lives*. London: Institute of Education Press.

Fogelman, K. (ed.) (1983). *Growing Up in Great Britain: Collected papers from the National Child Development Study*. London: Macmillan.

Hansen, K., Joshi, H. and Verropoulou, G. (2006). Childcare and mothers' employment: approaching the new millennium. *National Institute Economic Review*, 195: 84–102.

Joint Committee of the Royal College of Obstetricians and Gynaecologists and the Population Investigation Committee (1948). *Maternity in Great Britain*. Oxford: Oxford University Press.

Kiernan, K. and Smith, K. (2003). Unmarried parenthood: new insights from the Millennium Cohort Study. *Population Trends*, (Winter): 23–33.

Kuh, D. and Ben Shlomo, Y. (eds) (2004). *A Life Course Approach to Chronic Disease Epidemiology*, 2nd edn. Oxford: Oxford University Press.

Kuh, D. and Hardy, R. (eds) (2002). *A Life Course Approach to Women's Health*. Oxford: Oxford University Press.

McCulloch, A. (2006). Variation in children's cognitive and behavioural adjustment between different types of place in the British National Child Development Study. *Social Science and Medicine*, 62(8): 1865.

Osborn, A. F., Butler, N. R. and Morris, A. C. (1984). *The Social Life of Britain's Five-year-olds: A Report of the Child Health and Education Study*. London: Routledge & Kegan Paul.

Plewis, I. (ed.) (2004). *Millennium Cohort Study First Survey: Technical Report on Samling*, 3rd edn. London: Institute of Education. www.cls.ioe.ac.uk/studies.asp?section=000100020 0010010

Plewis, I., Calderwood, L., Hawkes, D. and Nathan, G. (2004). Changes in the NCDS and BCS70 populations and samples over time. Version 1. www.cls.ioe.ac.uk/library.asp?section=000100010 00600060018

Power, C. and Elliott, J. (2005). Cohort profile: 1958 British Birth Cohort (National Child Development Study). *International Journal of Epidemiology*.

Shepherd. P, Smith, K., Joshi, H., Dex, S. et al. (2004). *Millennium Cohort First Survey: A Guide to the SPSS Dataset*, 3rd edn. CLS, London: Institute of Education. http://www.cls.ioe.ac.uk/library.asp?section= 0001 0001000600060011

Wadsworth, M. E. J. (1991). *The Imprint of Time: Childhood, History and Adult Life*. Oxford: Oxford University Press.

Wadsworth, M. E. J., Butterworth, S. L., Hardy, R. J., Kuh, D. J., Richards, M., Langenberg, C., Hilder, W. S. and Connor, M. (2003). The life course prospective design: An example of benefits and problems associated with study longevity. *Social Science and Medicine*, 57: 2193–2205.

Chapter 6

Retrospective longitudinal research: the German Life History Study

Karl Ulrich Mayer

1 Introduction and overview

This article provides a comprehensive overview of the German Life History Study (GLHS). The GLHS is an almost entirely quantitative study based on nationally representative samples of eight birth cohorts in West Germany born between 1919 and 1971 and five birth cohorts in East Germany born between 1929 and 1971. Data was collected in nine different surveys between 1981 and 2005. These surveys were based on retrospective measurements in the sense that respondents were asked to recall episodes and events of their lives from the day of birth (e.g., place of residence) up to the time of the interview. Longitudinal data was recorded as event sequences in multiple life-domains and dated monthly. In this manner, time-continuous data covering all of past lives is being reconstructed. Altogether, more than 13,921 quantitative life histories in the form of multiple life domain event histories were collected from 11,441 respondents. At the time of the interview, the birth cohorts covered ranged in ages from 27 to 65. In addition, a component of the GLHS was incorporated in the Berlin Aging Study (Mayer and Baltes, 1996) where besides medical and cognitive assessments retrospective life histories were collected for 516 respondents living in West Berlin in the nineties between the ages of 70 and 103. Furthermore, a series of smaller methodological studies was conducted to assess and improve response rates and retrospective measurement.

A combination of six characteristics makes the German Life History Study distinctive:

1. It is a study of nationally representative birth cohorts.
2. It obtains longitudinal data by retrospective measurement.
3. Across cohorts it spans an historical period of more than 50 years and it is, therefore, a unique instrument for studying social change.
4. It covers several life-domains, especially both family and work.
5. By sampling both West Germany and East Germany it has an inbuilt design to study differences between sociopolitical systems and one important case of post-Socialist transformation.
6. It has invested more than any comparable study in the assessment of the quality of retrospective measurement as well as in careful data editing.

The German Life History Study started as an effort to trace changes in stratification processes and their embedding in contexts of social and economic discontinuity. It developed into a comprehensive research program on social mobility, life-course dynamics, transitions to adulthood, gender and cohort disparities, as well as the consequences of welfare state policies on patterns of life courses. With the opening of the Berlin Wall and the integration of the former German Democratic Republic into the Federal Republic of Germany, we extended the study to East Germany in order to reconstruct the former Socialist GDR society as well as the impact of the post-Socialist transition on life courses and the unification processes between the two Germanies.

The article is structured into five major sections. In Section 2 we will briefly summarize the motives and origins of the GLHS, its analytical goals, and the institutional contexts within which this research program was located. In Section 3 we provide basic information on the various surveys and their contents, as well as on the methodologies employed. Since the GLHS has been chosen for this handbook, among else, as an exemplar of retrospective longitudinal studies, we will focus in Section 4 especially on the issues, problems and solutions related to retrospective measurement. In Section 5 we will address the substantive areas for which the GLHS was intended, using and highlighting major findings from the study. In Section 6 we provide practical information on data access, data documentation, and give an introduction to the publications which have resulted from the GLHS.

We concentrate here on the more methodological and technical aspects of the German Life History Study and on a series of detailed study documentations (see 4.1) as well as on earlier overviews by Brückner and Mayer (1998); Diewald et al. (2006: Appendix 1); Hillmert and Mayer (2004): Ch. 11; Solga (1996); and Wagner (1996). For a general introduction into the research program of life-course theory and analysis, see Mayer (1990, 2000, 2004; Mayer and Huinink, 1990); for an explication of its theoretical rationale, see Mayer (2001); Mayer and Müller (1986).

2 Origins, goals and institutional contexts of the German Life History Study

2.1 Origins

The GLHS was initiated by Karl Ulrich Mayer and Walter Müller in the late seventies at the University of Mannheim (Germany). It grew out of four related research contexts. First, based on our prior work on intergenerational social mobility, we were interested in unraveling the mechanisms and processes generating socioeconomic inequalities across the life course going beyond the static comparisons in mobility transition matrices and the highly reduced structural model of status attainment research. Second, also in prior research based on census data, we had uncovered large differences between cohorts due to the large historical discontinuities resulting from World War II and the immediate postwar period. While the war and postwar turbulences were succeeded by the German “economic miracle” of the sixties until the mid-seventies by rapid educational expansion and occupational upgrading, it was unclear whether positive trends would continue or be reversed. Therefore, we wanted both to detail these cohort differences and extend and update them to more recent birth cohorts. Third, the GLHS project was part of a research group comprising both economists and sociologists from the universities of Mannheim and Frankfurt interested in issues of social and economic accounting, the impact of social policies, and the use of individual and household microdata for these purposes. Fourth, our primary prior data sources were individual-level longitudinal data from the Microcensus-Supplementary Survey from 1971 born between 1920 and 1940 (comprising 1% of the population, i.e., about half a million cases). This data source dried up because rigid data protection laws and rules were enacted and the West German Census Office lost interest in longitudinal data. Therefore, we were pushed to collect our own data rather than use secondary material.

2.2 Goals

The initial analytical goals for the GLHS derived from our sociological interests in social stratification, gender inequality, social change, and the welfare state. In regard to social stratification, we wanted to gain a better causal understanding of the mechanisms and processes generating socioeconomic and gender inequalities across the life course going beyond crude models of social background and educational attainment by taking into account more fully family histories, residential histories, educational and training trajectories, labor market processes, and family formation. In regard to social change, we wanted to employ cohort comparisons to trace more precisely social continuities and discontinuities as well as the impact of period and cohort effects. In regard to the welfare state, we were interested in the social consequences of institutional frameworks and social policies for patterns of life courses and life chances in the particular context of the "German social model" and its ongoing reforms. Methodologically, we were clearly fascinated by the new potential offered by micro-level longitudinal data, enhanced computing capacity, and recently developed dynamic modeling for exploratory, descriptive and causal research, as well as by the promise to transform biographical case studies to nationally representative samples.

2.3 Organization, institutional contexts and funding

In 1979 the GLHS started as a project of the Special Research Unit on Microanalytic Foundations of Social Policy (Sfb 3) of the German National Science Foundation at the universities of Mannheim and Frankfurt with Karl Ulrich Mayer as principal investigator. Support from this grant extended into the early nineties. An important boost for the design and early data collection of the GLHS came from the fact that in 1979 I also became a Program and later Executive Director at ZUMA, the German National Survey Research Center in Mannheim. Significant methodological solutions from the professional staff of ZUMA were provided on sampling, questionnaire design and fieldwork, coding, as well as data organization. It was a serendipity and crucial precondition for the further development of the study that in 1983 I became director at the Max Planck Institute for Education and Human Development in Berlin and Head of its Center for Sociology and the Study of the Life Course, an institute devoted to basic research. The Max Planck Association for the Advancement of Science provided through this Institute until the year 2005 the almost exclusive funding for the later cohort studies, the personnel resources, and a rich intellectual environment. In 2003 I accepted an offer from Yale University and the GLHS moved from Berlin to New Haven and my Center for Research on Inequalities and the Life Course (CIQLE). One of the cohort studies (on the 1964 and 1971 birth cohorts for West Germany) was cofunded by the German Labor Market Research Institute (IAB) and partly financed by the European Social Fund.

The German Life History Study is above all a collective effort of highly dedicated research and technical staff. In its lifetime of more than a quarter of a century, 18 full-time researchers have been associated with the GLHS, 16 professional and technical staff persons, 16 doctoral students, and more than ten post-doctoral students. Some of the senior research associates served at various times as study directors for various cohort surveys (among them Johannes Huinink, Martin Diewald, Britta Matthes, Steffen Hillmert, Heike Trappe, Götz Rohwer, Reinhard Nuthmann, Michael Wagner, and Ineke Maas). Erika Brückner served as the head of survey operations during crucial years of the study. Apart from the panel study on the 1971 cohort, all of the data collections were carried out in collaboration with commercial survey research firms. In temporal sequence these were GETAS in Bremen (Barbara von Harder), Infratest Social Research in Munich (Klaus Kortmann), and INFAS in Bonn-Bad Godesberg (Doris Hess).

3 Surveys and methods: sampling, data collection, data editing

Pilot studies for the GLHS began in 1979. From 1981 to 2005 we conducted five surveys (covering 8 different cohorts viz. cohort groups) and one panel study on West German samples and two surveys (covering five different cohorts viz. cohort groups) and two panels on East German samples (Table 6.1). For the purpose of this article we denote by the term "panel study" a design where some years later we re-interviewed the respondents of an earlier retrospective study and measured the interim period by continuous retrospective data.

3.1 The West German studies

The first West German component of the GLHS—collected in 1981–1983—constructed representative samples of three different groups of birth cohorts: 1929–31, 1939–41 and 1949–51 (Mayer and Brückner, 1989) with an overall sample size of 2171. This 1981–83 survey set in many respects the exemplar for the following surveys of the GLHS. It concentrated on small ranges of birth cohorts in order to capture fine-grained period and cohort effects. Basically, sampling costs prohibited focusing on 1-year birth cohorts and, thus, three-year bands served as the best compromise. Further, it established the basic recipe for data collection by focusing on retrospective event histories in separate life-domains (residence, family of origin, education, training, employment and careers). Events and transitions were recorded forward in time and dated monthly.

In the years 1985–87 the birth cohort of 1919–21 (Brückner, 1993) was added in two separate surveys (n = 1412), one by means of personal interviews and one by telephone interviews. By switching from personal to telephone interviews we showed that even very long standardized life histories (median 2–2.5 hours up to 6 hours) could be collected by using the telephone. Couples' decisions on retirement and the gendered impacts of old-age insurance policies became the focus of the analyses of this study (Allmendinger et al., 1993). In 1988–89 data for the cohorts born 1954–56 and 1959–61 (Brückner and Mayer, 1995) were collected (n = 2008). In 1997 and 1998 we extended the cohort series by collecting by telephone interviews almost 3000 life histories from the cohorts born between 1964 and 1971. Why this extension in regard to birth cohorts? It became clear very soon that by focusing on the cohorts born between 1930 and 1950 a very particular "success story" of continuous collective advancement would have to be written, while we had good reasons to assume that the 1930 cohort was much worse off than the cohorts born before and that a similar reversal was argued for the cohorts born after 1950. Also, we found it very attractive to include the earlier cohort born around 1921 because we could trace it up to retirement age and look at the impact of their war experiences. For the more recent cohorts, we knew that their demographic behavior had radically changed and we wanted to pursue explanations for these changes (Huinink and Mayer, 1995a). The 1964 cohort warranted special consideration because it was not only the largest in absolute size at birth, but grew by about a fifth through immigration. Both the 1964 and 1971 cohorts were of particular interest due to the economic downswings in the eighties and nineties and the alleged impacts of international competitive pressures. This interplay of cohort size, labor market conditions and policy measures became

Table 6.1 Cohorts and panels in the German life history studies

THEMATIC FOCUS	*Life courses and societal development*	*The war generation and the transition to retirement*		*Lost generation? Career entry in the labor market crisis*	*The impact of the baby boom: education, training, and early work lives*	*Early careers and family formation*	*Life courses and historical change in the German Democratic Republic*	*East German life courses after unification*	
Population	West Germany, including West Berlin					East and West Germany	East Germany, including East Berlin		
N	2171	407	1005	2008	2909	1073	2331	610	1407 (Panel)
Birth cohorts (age observed)	1929–31 (50) 1939–41 (40) 1949–51 (30)	1919–21 (65)	1919–21 (67)	1954–56 (33) 1959–61 (28)	1964 (35) 1971 (27)	1971 (34)	1929–31 (60) 1939–41 (50) 1951–53 (40) 1959–61 (30)	1971 (27)	1929–31 (65) 1939–41 (55) 1951–53 (45) 1959–61 (35)
Response rate	62.3%	48.8%	73.3%	86.1%	66.1%	N/A	52.2%	49.5%	74.1%
Field period	1981–83	1985–86	1987–88	1988–89	1998–99	2005	1991–92	1996–98	1996–97
Data collection mode	Personal interviews		Computer-assisted telephone interviews		Computer-assisted telephone and personal interviews		Personal interviews	Computer-assisted telephone and personal interviews	
Core contents	Detailed event histories: family formation and fertility, education and training, employment and interruptions, residential mobility. Detailed questions on family of origin, including siblings. All studies also include some questions on politics and religion, current economic situation, and on the interview situation.								
Specific contents	Detailed work histories for spouses; assets and savings; activities, interests, needs	Impact of wartime events on respondents' life histories; political socialization; transition to retirement		Additional questions on vocational/ professional training, labor market entry, future aspirations	Labor contracts and job decisions; occupational control beliefs; memberships (politics, associations, religion); household structure and income, life satisfaction	Impact of delayed career and family formation; work–family issues	Economic situation; changes and experiences before and after fall of the wall; social networks and informal exchange	Membership in organizations and associations; party preferences; social support networks; control beliefs and various other psychological scales; economic situation	

the central focus of our monograph on the latter two cohorts (Hillmert and Mayer, 2004).

3.2 The East German studies

When the Berlin Wall fell, we seized the opportunity to extend the study to East Germany and its transformation because this extraordinary opportunity presented us with an exemplary case, a natural experiment, for studying life courses under the conditions of extreme societal discontinuities. Data on four cohort groups were collected in East Germany in 1991–92: 1929–31, 1939–41, 1952–54, and 1959–61 (n = 2331) (Goedicke et al., 2004). We had to rely on personal interviews since telephone coverage was still low in the East. The selection of the specific cohorts was partially due to our attempt to match the West German samples and to adjust for specifics of East German sociopolitical history. We went back to a fraction of our East German respondents with a mailed questionnaire in 1993 (enlarging on personality and network variables) (n = 610) and interviewed them on their intervening life trajectories again in 1996–97 (n = 1407). Our primary goal with this panel was to cover more of the transformation process and to explain its outcome in a life-course framework. At that time we added the 1971 cohort (n = 1407) for East Germany which was then re-interviewed in 2005. We used these studies on the one hand for a reconstruction of lives under the Communist regime in the GDR (Huinink and Mayer, 1995b) and for a comprehensive study of life courses during the post-Socialist transformation (Diewald et al., 2006).

3.3 The 2005 panel study and qualitative biographical complement

In the year 2005 we re-interviewed 1073 of 1805 of the men and women born in 1971 from both the 1997 East German Study and the 1998 West German Study. We had four motives for the panel study. First, when we interviewed the 1971 respondents they were about 27 years old. That means few of them were married, almost none of them had children, and about a quarter were still in training. In order to capture their full transition to adulthood and to unravel the mysteries of their delayed family formation, we were highly interested to follow up on their life trajectories. Second, we had developed our own computer-assisted telephone life-history questionnaire, especially suited to optimize recall, and wanted to test this instrument under normal survey conditions. Third, we had for the first time the opportunity to conduct the study in our own telephone interviewing laboratory and, thus, to exert full control over the process. Finally, we wanted at last to combine quantitative and qualitative methods of data collection. On the basis of the quantitative protocols we selected a small sample stratified according to gender, East–West, North–South, Urban–Rural and High–Low qualifications. In cooperation with the Berlin Institute of Social Research we conducted 27 narrative biographical interviews which are available both on tape or digital record and in transcript. Field time ran from early 2005 to the end of June 2005 and was truncated due to restricted funds. Selectivity should, therefore, be carefully monitored on the basis of the initial surveys.

3.4 Methods studies

In addition to the nine major surveys, we conducted a series of supplementary methods studies. This started in 1979 with pilot studies testing for the accuracy of recall based on a local cohort survey carried out 10 years before and testing for variants of questionnaire design (fully standardized domain-specific event histories versus partially structured narrative interviews versus life-history calendars). In the context of the 1997 panel study of East Germans, we carried out a nonresponse study (Wehner, 2002). Moreover, for each of the studies assessments of representativeness were performed using relevant cross-sectional census data (e.g., Blossfeld, 1987b). Several studies on the reliability of retrospective measurements were also conducted (Brückner, 1995; Reimer and Matthes, 2006).

3.5 Sampling

Our decision to use birth cohorts as population units instead of the entire population was primarily motivated by prior work about cohort differences in Germany based on census data. Financial restrictions ruled out obtaining a sample covering the full population and, at the same time, gaining a sufficient number of people in single birth cohorts or narrow birth cohort groups. Therefore, specific cohort groups were selected. Our prior work with census data provided the criteria for these cohort selections. For instance, the 1939–41 cohorts were chosen because they were at the peak of the greatest baby boom of the century. The 1929–31 cohorts were selected because we already knew that these were the cohorts most affected by World War II and the breakdown after the war. The cohorts 1954–56 (in West Germany) and 1952–54 (in East Germany) were chosen due to their pivotal role in demographic changes (on the selection of cohorts, see also Mayer and Huinink, 1990). The 1964 cohort was selected due its exemplary nature as the most recent baby boom cohort. Given overall sample sizes well below 40,000, there are good reasons to rely on cohorts rather than cross-sections in longitudinal surveys (Featherman, 1979), but there are even more cogent reasons not to rely on one single cohort in either retrospective or prospective studies.

Thus, the sample design required selecting people born in specific years. In contrast to representative, cross-sectional samples of a population, the cohort-centered selection of samples (with a sufficient number of people in each cohort) requires special procedures.

This section deals, in particular, with the practical consequences of using a cohort-centered approach in a nationwide face-to-face and telephone survey. Arriving at a representative sample made up of different adjacent and nonadjacent birth cohorts requires extensive preparatory work. Since the sample is limited to only these cohorts, it had to be selected out of the entire corresponding population. Four different methods were used to accomplish this. For the face-to-face surveys (1981–83 and 1985), household listings based on the ADM Master Sample were collected to identify the target population. On the basis of these 13,974 private households, a sample of cohort members was chosen.

The telephone surveys were based on an initial large-scale representative sample of households with telephones, out of which the informants for the second survey of elder cohorts and the two samples for the younger cohorts were selected. Starting with the 1954–56 cohort, we relied on the now electronically available residential registers for pre-selected counties and cities. For the East German samples the totalitarian character of the GDR regime proved helpful for the sampling design. The GDR had a unitary population register which could directly be used to draw regionally stratified samples of members of certain birth cohorts.

Special strategies were necessary to compensate for specific, regional losses, which affected primarily major cities or urban centers in the first field survey and primarily the points lying in rural areas in the second. In both of these (face-to-face) surveys, it was necessary to make repeated follow-up trips in order to interview informants who were difficult to reach geographically or who were indecisive about participating in the survey. Besides the geographical dispersion of the informants, another probable reason for the slow tempo at which data was collected was the strict procedure used to select the informants (no substitutions could be made on either the household or the individual level). Response rates of around 60% may raise doubts about whether the results obtained in laborious field surveys were worth the effort required to obtain them. However, these response rates are close to the ones for the usual cross-sectional surveys in Germany despite the fact that our demands on the respondents' time were much higher.

Switching to collecting data by telephone does not only allow better supervision and selection of interviewers, but also made a more centralized field operation possible and, likewise, appeared to present a viable solution in identifying samples of specific cohorts. The rate of those not at home could be reduced from nearly 4% to 1.5%. Furthermore, the "direct line" to the respondents further made it possible to split up the long interviews into several telephone conversations. This proved to be a big help, especially for the oldest cohort, whose interviews were sometimes extremely long (one-third of all interviews in the survey of the eldest cohort were conducted in two or three calls). In view of the mobility of the younger cohorts, the up-to-dateness of the telephone sample, as well as the ease of repeated contacts, proved to be an advantage in the process of data collection.

3.6 Methods of data collection

Life-domain specific event histories, PAPI and CATI-instruments

There were only a few examples of quantitative life-course studies using large samples when the GLHS was initiated. In the mid-sixties a representative survey of the male population known as the Johns Hopkins Study was carried out by Coleman and Rossi in the United States in 1968, and a similar survey, the Norwegian Life History Study, was carried out by Rogoff Ramsoy in 1971. Allmendinger (1989) used these Norwegian and US data together with the GLHS for a study on educational stratification and career processes. Noteworthy studies originating after the GLHS and mostly modeled on it are the Swedish Life History Study (Jonsson and Mills, 2002), a Dutch Study (de Graaf, 1987), and the Swiss Life History Study by Marlis Buchmann (see Buchmann et al., 2006), and also the study conducted by Mach (2003) which is a direct replication of our 1971 cohort study.

The German Life History Study differed from earlier studies in two important aspects. First, it is more representative: including women and a broader range of birth cohorts. Second, it is more comprehensive with regard to life-domains: including an education, training and employment history, a full residential history not only of locations but also of apartments, a family history, and a number of other important thematic areas of the life course such as parents and children. The goal of the questionnaire construction was to represent the "natural history" of individuals on society to the greatest extent to which it could be rendered strictly comparable between subjects and be made quantifiable.

We also parted ways with earlier (and later) studies by not employing a life-history calendar. In a life-history calendar yearly or monthly time defines the rows of a matrix, whereas types of events (e.g., schooling, employment, marriage, child birth) define the columns. The advantage of a life-history calendar is that all events and the time dimension are represented on a single sheet of paper (or electronic equivalent), thus facilitating comparability and consistency across domains. The big disadvantage of life-history calendars lies in the fact that only very little information can be inserted on specific episodes. This format also implies a rectangular data array where all months or years must contain some information. The advantages and disadvantages of multiple-domain event sequences are complementary. This format is efficient in data storage, since it focuses on a sequence of spells or episodes and only needs to list beginning and ending dates. Thus, it corresponds both to the way the information is actually solicited in an interview and the way it is used in survival and event history analysis. Most importantly, it allows us to collect a relatively large amount of information for each given episode. For instance, for any given job we collected data on employment status and occupational title, industrial sector and size of firm, hours worked, type of contract, beginning and ending income, as well as reasons for job changes. Concentrating on specific life-domains actually facilitates recall, but requires more investment in checking on inconsistencies across life-domains. In the earlier (paper-and-pencil) interviews this was either accomplished by fold-out pages or by data editing after the interview. In the CATI-versions these checks were done automatically leading to additional questions on potential inconsistencies.

Thus, in developing the questionnaire we first reviewed life-domains: family of origin and one's own family history; education and training; residence and household; employment, income, and consumption; social, religious, and political participation; friends and informal networks (in the first survey also disabilities and the medical history). For each of these life-domains, we first tried to convert it into continuous event and state histories. We started with the idea that for each life-domain there is a more formal, institutionalized, and legitimate path, as well as a more informal, marginal, and less institutionalized path. Thus, besides natural parents, we asked for foster and step parents. Besides (full-time) formal schooling and training, we also asked for part-time and for further education and training. Besides main jobs, we also asked for additional jobs and marginal employment. And in regard to family formation, survey by survey we extended the marriage history to a history of cohabitation and of partners. Limitations of the already excessive interviewing time as well as measurement problems led us to cut down the life-domains which were finally included. For instance, we did not use an instrument for diachronic associational membership or for friends across the life course. Also, we economized on the household composition history by attaching—in some of the surveys—a shortened version of the residential history (only locations). We also dropped a good and tested instrument for consumption history, namely a history of bought cars and their properties.

Questionnaire design

The following factors had to be taken into account while designing the questionnaire:

1. The degree of complexity of the questionnaire, which had to record a great deal of information and also do justice to the interindividual, group-specific, and cohort-specific variance of life courses.
2. The historical conditions under which structurally equivalent life events took place in different historical periods.
3. The recording of time (frequencies, durations, and absolute and relative points in time) both as a simple measurement (dimension) and as a way of structuring the life course.
4. The "sensitive" topics in the respondent's life and the socio-demographic data and life events of the respondent's relatives (parents, siblings, spouse, children, and grandchildren).
5. The adaptation of the questionnaire to the size and field conditions of a representative national survey, i.e., the quality of the interviewers.
6. The specifically retrospective character of the data collected (Brückner and Mayer, 1998: 161).

Cohort-specific questionnaires

In one sense, cohort-specific questionnaires are merely a specialized case of screening, e.g., if the respondent belongs to this cohort, then ask the following set of questions. However, they go beyond mere screening in that the actual content of the questionnaire is changed. German institutions and society as a whole have undergone a great many changes in the past hundred years, and people growing up in different historical periods have encountered widely varying historical conditions. To name just one obvious example, the older cohorts experienced the war and its immediate aftermath, while the younger ones did not. Similarly, the National Socialist educational system in which the oldest cohort received its schooling differs considerably from the West German educational system after World War II, which itself has undergone several fundamental changes over the last forty years.

These differences had to be taken into account in the questionnaire. The phrasing of certain questions as well as the meaning or the values of certain variables thus depended on the cohort which they describe. Although this approach usually limits the extent to which different cohorts can be compared, this was deliberately ignored in favor of a more accurate reflection of historical reality.

Use of computer-assisted telephone interviewing

The first two surveys were conducted using the classical personal or face-to-face interviewing method. This method led to several difficulties, some of which will be discussed below. Due to these difficulties, the decision was made to conduct further surveys via telephone. Once this decision was made, it was but a small step to replace the previously used paper questionnaire with a computerized questionnaire; in other words, to make use of computer-assisted telephone interviewing or CATI. This proved to be an extremely beneficial decision. One of the most important advantages of using CATI was the ability to automate the screening process. The interviewer no longer had to leaf through a long questionnaire filled with little arrows and boxes with screening instructions. Instead, the computer would automatically screen questions and display the next question to be asked on a CRT screen. This freed the interviewer to concentrate exclusively on conducting the interview and recording the data as precisely and accurately as possible.

Moreover, an unlimited number of data validity and consistency checks could be automated. If erroneous or inconsistent data were entered, the interviewer would immediately receive a message asking for a correction or confirmation of the data. Data validity was checked merely by comparing the data entered with previously established valid values or ranges.

Data consistency was checked by comparing the data entered both with previously entered data and with previously established plausible ranges of values. For instance, it might be assumed that women have their first child between the ages of 18 and 40. By comparing a woman's year of birth with that of her first child, her age at the child's birth could be determined. If it fell outside the previously established range of plausible ages, a message asking for confirmation would be displayed, e.g., "Was your mother really 14 years old when she had her first child?" Thus, a major source of errors could be eliminated, i.e., corrections could be shifted from the later process of data editing and high cost re-checking with the respondent to the initial interview.

Additionally, questions could incorporate information gained from previously asked questions and could, thus, be made very precise and specific; such as, "Until when did you work as a file clerk in the bookkeeping department of company X?" Besides the fact that such questions vastly reduce the odds that the interviewer or respondent will confuse one event with another, such questions also give the respondents the impression that the interviewer is an intelligent person paying attention to what he or she is being told.

One of the few disadvantages of using CATI was the loss of transparency offered by a large matrix on a single sheet of paper, e.g., per a given life-domain. Due to the physical constraints of a computer system, namely only a given amount of information can be displayed on a screen at one given time, it was necessary to split the matrices up into smaller templates, i.e., sets of questions which could fit onto one screen. However, the loss of transparency was more than compensated for by the automated consistency checks described above. Furthermore, it was possible to maintain an overview of the data by switching between displays of information at different levels; for instance, one display showing a list of all the jobs an individual had held and another display showing detailed information for only one of those jobs. However, we did observe that telephone interviewers sometimes took notes on a piece of paper to keep track of interdomain consistency. Technically, these problems could easily be remedied by larger and split screen monitors.

Fieldwork

Problems with conducting the interviews were expected from the outset for several reasons. To begin with, the format of the questionnaires appeared unusual and unfamiliar to the interviewers. It was unknown to what extent the respondents would cooperate, given that they would be asked so many questions of a personal nature. Moreover, the survey firms entrusted with the fieldwork had to deal with tasks much more demanding than those they routinely encountered. In general, the difficulties encountered during fieldwork are attributed to a lack of cooperation on the part of the informants. However, this did not prove to be true for the study described here. On the contrary, with few exceptions, the respondents exhibited an almost astonishingly positive reaction to the interview. It was, in fact, the interviewers who presented the most difficulties throughout the entire data collection process (even after the introduction of CATI), although a great deal of effort was made to supervise and support them.

3.7 Data editing

Editing plays an extremely important role in the processing of life-course data. The first and probably major task is to record the data provided by the respondent in formats which are consistent. For instance, if respondents listed marginal jobs when asked about their main job history they had to be transferred into the data array for second and marginal jobs. Second, the sequence of chronological events and numerous events linked to one another in terms of content had to be combined into a single, individual life history free of inconsistencies within and between life-domains. Completeness and plausibility are important criteria, both on the individual and the interindividual level (for a detailed description of the process, see Brückner, 1993; Hillmert, 2002). Consistency and plausibility as well as the continuity of sequential events can be checked in detail by using the intrinsic logic of the events and their relation to one another as well as to institutional and historical contexts. In this respect, editing serves as a sort of "internal" validation. Like questionnaires, editing methods have to be adapted to the specific historical context of each cohort.

The treatment of missing values was a particular problem. Completeness is an important prerequisite for the analysis of sets of event data. Gaps in the data had to somehow be filled, even if follow-up research (which included going back to documents or respondents) could not produce exact information to fill them. Dates were reconstructed by using so-called artificial months; for instance, 21 for January = beginning of year, 26 or 27 for June and July = middle of the year, and so on. Landmarks, which are usually major historical events used to set a relative date for personal events, also played a role in the editing. These relative dates ("... it took place in the same month as the assassination attempt on Hitler ...") occasionally misled the interviewers into entering false dates, despite the fact that they were provided with a chronological list of historical events. Using history books and similar reference material, the editors were later able to reconstruct the actual dates on the basis of the respondents' relative dates, which were usually authentic and very accurate (they are referred to as "flashbulb" phenomena in psychology). Residential and employment history were compared to check for data validity and consistency and also to enhance or clean the data in difficult cases.

We relied on several sources for data editing:

- the paper-and-pencil or electronic records
- biographic abstracts produced from the data
- the tapes which we recorded from all the respondents who volunteered (most of them did)
- written inquiries going back to respondents
- telephone inquiries going back to respondents.

Across the surveys we contacted about a fifth of the respondents again personally in order to resolve real or potential errors.

An ever-increasing number of data validity and consistency checks were computerized during the course of the study, but computers still could neither completely eliminate nor even significantly reduce the task of editing the data by hand. All in all, the amount of time and money needed for the data to be edited was about the same or even more than that needed for the interviews themselves.

4 Autobiographical memory and retrospective measurement

Although data collection in a retrospective life-course study has a cross-sectional character (a population sample is interviewed only once at a given time), the data itself corresponds, in principle, in content to that of a sufficiently long prospective panel study, including information on the entirety of the respondents' lives. The caveat "in principle" relates to prior selectivity (emigration or true mortality) in retrospective studies and panel attrition in prospective studies.

The use of retrospective data collected in a "one-shot" survey offers a plausible alternative to prospective longitudinal data collection, especially in the case of life-course studies (see Featherman, 1979; Solga, 2001). It is time-efficient and cost-efficient. It is time-efficient because one does not have to wait for a long time until prospective data spans a time length of sufficient duration to be of scientific interest. It is cost-efficient because longitudinal data can be obtained by one or a few interviewing phases and survey organization does not have to be maintained over a long time. Retrospective data collections are often also preferred to prospective panels because both the quality of survey instruments and the scientific interest in the subject matters tend to become obsolete in panels which run for a longer time.

Objections raised against retrospective data, in particular, regarding the accuracy and precision of people's memory, apply to all kinds of biographical data. Even cross-sectional studies often contain questions of retrospective character, usually dealing either with the frequency or duration of certain events such as schooling or medical treatment or with the time at which a given event took place, e.g., age at marriage or divorce, birth of children, or employment dates. Most frequently such retrospective questions in cross-sectional studies are of a much more ad hoc nature than if they are asked in the context of event sequences. They are, therefore, more prone to recall biases. Also, most prospective panel studies, in fact, collect some retrospective information to cover events and changes between the points of data collection and to compensate the arbitrary starting points of a panel in the lifetime of its sample members. It has been estimated that in any given wave only a third of the questionnaire content actually refers to concurrent states. It is also worth mentioning here that major prospective panel studies, like the US Panel Study of Income Dynamics (PSID) or the British Birth Cohort Studies, do not actually allow construction of continuous event histories. If they allow continuous data, such as job sequences, e.g., by reconstructing what happened across the past year by a monthly calendar, they often suffer from inconsistencies between recalled and concurrent data. Most significantly, prospective panel studies rarely provide representative samples of life histories due to attrition or perforated coverage. Solga (2001) has estimated for the German Socio-Economic Household Panel that between 1984 and the end of 1998 only 25% of the life courses of the respondents in the initial sample were complete. The advantages and disadvantages of retrospective versus prospective panel data can, therefore, not simply adjudicate in favor of prospective studies. It might appear that the problems of both retrospective and prospective survey data could be solved by the increasing (and important) use of longitudinal register data. Apart from still rampant problems of access, register data are far from limitations. Often administrative purposes in variable measurement do not match scientific objections, variable definitions change over time and certain series only start at recent historical times. Thus, especially if one wants to go back in time there is often no alternative to retrospective designs.

That said, it is no doubt that potential problems of retrospective measurement constitute the biggest challenge to the use of such studies in collecting individual-level longitudinal data. Therefore, we devote the following sections to an assessment of how much error one has to expect in using retrospective data and whether survey instruments can be fine-tuned to reduce such error. Obviously, to hide one's head in the sand, like the proverbial ostrich, will not help. Rather one has to tackle the issue head on from the very beginning of designing survey procedures and especially questionnaire and fieldwork. My general thesis is that the quality of instrument design, fieldwork, and data editing is much more salient for the reliability of the data than the question whether it is retrospective or concurrent. Recall errors are only a small fraction of the total survey error (Groves, 1989). At the outset, let me make two observations. First, it is obvious that for certain kinds of variables concurrent data is almost always better than retrospective data. This is, for instance, the case if one is interested in household data (and wants to interview all household members) or if one wants to collect detailed data on income or on attitudes. I stress "almost always" because we found that the self-employed report their incomes retrospectively often much more truthfully than concurrently (for fear of the Internal Revenue), and traumatic events like marital separation are often revealed better with some time distance to the event. In a ten-year follow-up we found even the subjective reasons for not being able to attain the desired occupation to be very highly stable. Second, we are today in a much better situation of dealing systematically with retrospective error than at the time when we started the GHLS. At that time very few reliability studies were available and were often of an ad hoc nature. The psychology of memory with its sole distinction between episodic and semantic memory was light years away from offering either adequate analytical or experimental results to be of any use. This distinction suggested that recall in autobiographical contexts might be even more error prone than recall in general. In the meantime, the psychology of autobiographical memory has rapidly advanced (Rubin 1986, 1996) and the emergence of cognitive survey psychology has made major progress (Schwarz and Sudman, 1996; Sudman and Bradburn, 1996).

Within our research group Maike Reimer has systematically developed the extent and kind of relevance of autobiographical memory retrospective measurement and has conducted several empirical studies to assess the range and types of retrospective error (Reimer, 2001, 2005a, 2005b; Reimer and Matthes, 2006). The psychology of memory has demonstrated that persons tend to reconstruct their own biography by selective forgetting and by a change in attached meanings in order to enhance subjective self-worth and identity. The question is whether similar effects occur in the measurement of most objective life-course data, like schooling, occupational trajectories, and family events. Are there types of data which are more or less reliable? Are recall errors a function of time distance between the event and the interview (which would suggest that retrospective measurement for small intervals like panel interviewing times are okay but long recall periods might not be). Are the errors which can be revealed consequential for the types of substantive conclusions we are interested in? Can survey instruments be fine-tuned to improve recall and to minimize recall errors?

The insights from the psychology of autobiographical memory cannot, however, be applied easily to standardized retrospective surveys. Quantitative social science measurement is above all selective in the sense that it is highly limited to types of data which are valid, reliable, standardized, complete, and comparable between subjects. In contrast, autobiographical memory functions, per se, as a selective, constructing and dynamic process which forms a broad pool of episodes and events. It selects, presents and encodes subjectively relevant ones which are then retrieved again selectively and in a biased manner. In this sense, the usual results of autobiographical memory research apply much more to qualitative biographical research than to standardized retrospective measurement.

Nonetheless, in both cases encoding in a temporary (working memory) and a long time storage "server", as well as the retrieval processes from these memories, are of crucial nature. When persons encode or retrieve episodes and events, these must be mobilized as representations. Autobiographical memory relates to past personal experiences and biographical facts without experiential depth. Instances of one's past are remembered better if they fit into more general schemata (Brewer, 1986). This is highly significant for sociological life-course research since it predominantly is interested in highly institutionalized and public trajectories rather than highly idiosyncratic private lives. Such schemata are also life phases and their sequences (Conway and Pleydell-Pearce, 2000). However, it is not entirely clear how the results of memory psychology on life stages relate to the spell sequence character of event histories. They seem to suggest that the normal and typical will be remembered more easily than the nonconventional and atypical. Although this seems plausible, it is not a logical consequence, since deviations from the normal could be encoded and retrieved on the basis of the more generic schema.

At any rate, retrospective measurement in life-course research is supported by the mechanisms of autobiographical memory. For both life-course structures, such as sequences and temporal relations, hierarchical orderings (e.g., main job and second jobs) as well as horizontal relationships, such as between employment and the family situation, are important as mechanisms of representation. Thus, it follows that there is an isomorphism between the manner of how life courses are structured in society and how autobiographical memory functions optimally. This is crucial in our context, since processes of remembering and retrospective measurement are not exclusively connected as sources of error (selectivity, biases, simplification), but much more importantly built upon the very same cognitive "grammar." Personal memories are organized through self-schemata, i.e., generalized expectations a person has about him- or herself. These self-schemata are closely connected with narrative representations of individual life histories. Therefore, remembrances which do not fit into such schemata are subject to simplification, and changes toward the conventional and smoothing.

For retrospective measurement it is of special interest how according to autobiographical memory temporal information is being stored and retrieved. Chronological patterns appear to be more salient for retrieval than for encoding. Moreover, time distance to the event is related to recall reliability in a curvilinear fashion. Recall is best for very short time distances, then worsens, but stays very stable as time distance goes on. This result is crucial, since it could support retrospective studies with a long time frame into the past. In contrast, errors in dating events seem to increase with time distance in a linear fashion. Experimental data shows that early childhood events are badly recalled, events for the age between 15 and 25 are well remembered, and little is recalled for middle age. This, however, suggests that recall is not only a function of time, but also a function of the probability of occurrence of given events.

What then can we conclude from the psychology of autobiographical memory for retrospective measurement in life-course research?: "Through selective forgetting and reinterpretation persons tend to construct consistency . . . . Personal memories are simplified and more coherent versions of the actual life history . . . and conform more to conventions and social expectations . . . negative experiences are more often forgotten or reinterpreted as positive ones . . . . Personal data can be recalled reliably and robustly even after a long time, if the representations match autobiographical memory, if they fit into self-schemata, if there is a rich relational structure, and if recall has been frequent and if biographical details can function as substantive or temporal cues" (Reimer, 2005a: 51 and 52; translation by author).

In regard to the total survey error (Groves, 1989), recall errors are a potential part of observational errors, such as social desirability. Recall errors can be of very different kinds. They might relate to incidence and frequency of episodes or activities (too few, too many), timing (ending, beginning), direction (too short, too long) and extent (e.g., monthly vs. yearly misdating), inconsistencies between life-domains (e.g., marriage without cohabitation), and differences between subgroups of respondents (e.g., all family events are better remembered by women) (Reimer, 2005a, Ch. 3). Recall errors, however, might also—if handled badly—lead to serious nonobservational errors as a consequence of sloppy sampling, data editing, and data organization. Thus, in the pragmatics of survey research recall errors must always be seen in the context of total survey error.

Prior research demonstrated recall errors of different kinds and magnitudes, but results are frequently inconsistent and almost exclusively only related to short time distances. It also remains unclear how much of observed discrepancies were due to flaws in field research or more fundamental problems resulting from faulty memory (Reimer, 2005a, Ch. 3). There are basically two research strategies available to uncover the salience of recall errors. The first strategy is short-term or long-term replications, i.e., the same respondents are being asked about the same time of their life repeatedly. The second strategy relates to systematic comparisons with evidence which is not channeled through the respondent, e.g., personnel documents or administrative data. In the GLHS we applied both strategies to assessment of the potential problems of retrospective measurement. On the one hand, in the East German Panel Study we asked respondents again in 1997 about their lives after 1989, a period which we already had covered partially in the 1991 and 1992 study. This overlap period thus served as a basis for the measuring of recall consistencies. On the other hand, we asked respondents to allow us access to their social security files which record all labor earnings and the full employment history (except for self-employed and civil servants).

Reimer shows that, among else, "salience" seems to play a large role in recall processes. While after five years only 89% name the same number of children, the error drops to 3% for legitimate children. A conceptual lack of clarity ("only my children or also the ones of the partner") and emotional closeness or distance (dead or adopted children) influences recall error. The error rate or, technically more adequate, the inconsistency rate is twice as high for men than for women. Incidence seems unaffected by time distance between event to be recalled and time of the interview, but the magnitude of deviations increases. In regard to the timing of leaving the parental home, in the second interview very early or very late departures were brought closer to the median. In regard to job histories, in the later interview the number of jobs was reduced (by about 10%) and some formerly split episodes were fused. Although especially episodes of short duration and of low frequency are often better recalled when the interviews are not very distant in time, for East Germans it was recalled more frequently in the five-year follow-up. This is clearly the case because unemployment was a totally new phenomenon for East Germans, it was conceptually unclear, but also individually shameful. Five years later, both of these conditions did not hold anymore, unemployment had become a very public and a collective phenomenon, and was, therefore, mentioned more often. As to timing (in months) very short deviations dominate the results and, thus, are not very consequential for substantive analyses. An interesting result emerged also in relation to the fall of the Wall. As an historical anchor it should have been especially useful for recall (Loftus and Marburger, 1983). In the German case, it, however, also led to some recall confusion, since German unification actually spanned between the fall of the Wall in November 1989 and formal state unification in October 1990. Since the fall of the Wall seems to be the stronger anchor, some respondents dated episodes of the fall of 1990 erroneously back to 1989. Altogether, the less conventional and more complex a life, the more likely there will be recall error. This is an important finding because it introduces a systematic bias into retrospective measurement, which one has to take into account in interpreting such data. The complement in prospective studies is that such respondents are more likely to drop out or cannot be reached. The more institutionalized, the better the recall. For instance, do respondents recall the duration of employment episodes better than of nonemployment episodes? A further important finding relates to the interrelatedness of errors and of reliability. Reimer could show that errors in timing are rare, once the type and sequence of episodes is properly recalled. This seems to corroborate a tenet of the psychology of autobiographical memory, i.e., that chronological calendars as such do not operate as the basis of recall and retrieval, but rather the substantive structure and sequences of episodes. This clearly gives some boost to the life-domain specific collection of life-course data in contrast to life-history calendars. The comparison with register data corroborates the recall risk of nonstandard episodes, like unemployment, which are markedly less well remembered (or probably revealed) than in register data. In our case it also demonstrated that in an employment register, family events were much more faulty than in the retrospective survey.

In the GLHS, various features of the data collection procedure provide beneficial support for the recall task: single events are recalled within thematic domains in forward chronological order, extensive opportunities for cross-references of memories are provided, complete sequences are collected which avoids boundary effects, calendar dates are reconstructed from the life-history context and partly anchored with a landmark event. Moreover, in every part of the study, methodological innovations were being developed, introduced and empirically tested, and opportunities to check data reliability and validity were used or intentionally created. This concerns various aspects of data collection (representativeness, nonresponse, interviewer effects, retrospective measurement, recall errors), data edition, and data linkage.

Nevertheless, an examination of reliability and external validity shows that in retrospect, respondents omit or insert, merge and dissect, temporally stretch or compress episodes, label episodes or transitions differently, and move transitions forward and backward in time. This leads to an overall reduction in complexity and change in the recalled life course. Observed discrepancies seem to be influenced by the individual and institutional life course and recall contexts in which an event occurs and is remembered, maybe more than by the time elapsed since an event. Such error can be greatly reduced by a meticulous and labor-extensive single case data edition. This, however, depends on which aspect of error we look at (Reimer, 2005b).

In the GLHS we have also introduced and evaluated new features into the data collection procedure that emphasize and strengthen the role of the interviewer as interface between social scientific concepts, data and standardization requirements, and memory processes. A special data collection instrument allows interviewers to act as “reconstruction guides” and effectively help respondents reconstruct episodes as intended and date them correctly and consistently. Preliminary results indicate that this reduces editing costs and improves data quality (Reimer and Matthes, 2006; for similar instrument development, see Belli, 1998; Belli et al., 1999; Belli et al., 2001).

In sum, one clearly cannot assume naively that retrospective data is free of recall error (see also Grotpeter, Chapter 7 in this volume), but, among other considerations, the drop of such errors after careful data editing shows that the quality of the survey process is the predominant factor in regard to data quality. The question is, therefore, not whether one type of data is error free and the other is not, but which different errors must be reckoned with in different kinds of longitudinal data.

5 Substantive areas and major findings

5.1 Core empirical areas

The core substantive areas which overlapped in all of the surveys of the GLHS were: residential history of locations, parental background, siblings (including their schooling), school trajectory, vocational and professional training, further education, employment and occupational history (including sector, firm size, labor income), marital history, fertility history and education of children, education and careers of spouses and partners, denominational history, current household income. Areas covered only in one or less than all of the surveys are residential history on the level of apartments, second jobs and marginal employment, nature of employment interruptions, history of cohabitation and partnerships, illnesses, control beliefs, subjective occupational mobility, history of marital satisfaction, life satisfaction, sociopolitical attitudes, types and intensity of social relationships.

5.2 Major findings and publications

Stratification and intergenerational social mobility

In accordance with the initial goals of the study investigations on class structure, stratification, social mobility and status attainment, their mechanisms and changes over time in the German context and in cross-national perspective were a major focus of the analyses based on GLHS data (Mayer and Aisenbrey, forthcoming; Mayer and Carroll, 1987; Mayer and Solga, 1994). In another aspect of this work the effects of educational expansion on inequality of educational and occupational opportunity were assessed (Henz and Maas, 1995). These studies corroborate on the one hand the pervasive structuring of socioeconomic inequalities in (West)

Germany, not least via a selective educational system, the persistence of the apprenticeship and academic exclusiveness, and on the other hand a weak trend of increasing social opportunities (which might have come to halt most recently).

The dynamics of careers

Intragenerational mobility was a special focus of the GLHS in regard to job shifts, firm shifts, and occupational shifts. Beyond confirming the general finding that such shifts decline with duration in the labor force, these studies provide ample evidence for the higher stability of German workers, their higher firm attachments, and the pervasive occupational structuring of careers. Increases in intragenerational shifts were more pronounced in the fifties and sixties than in the eighties and nineties (Allmendinger, 1989; Carroll and Mayer, 1986). Marked cohort effects, e.g., for the 1964 baby boom cohort, could also be demonstrated which, however, were partially offset by political measures (Hillmert and Mayer, 2004).

Education, training and the early career

Despite massive educational upgrading, the transition to the labor market proves to be remarkably robust during the 50 years which are spanned by the GLHS cohorts, and few effects of educational inflation and mismatches could be detected (Blossfeld, 1987a; Pollmann-Schult and Mayer, 2004). Participation in further education gradually increased across cohorts, but concentrated in a few years early in the careers. It also exacerbated rather than diminished educational inequalities across the life course (Becker and Schömann, 1996). However, the transition phase between the end of schooling and labor market integration became more checkered with more interruptions and multiple-training episodes (Jacob, 2004). Fixed-term contracts became more widespread at career entry, but are mostly beneficial for further advances. Huge problems have developed for unskilled youth (Solga, 2005).

Residential migration, leaving home, family formation and dissolution

The cohort studies of the GLHS cover marked discontinuities in family formation. War-related delays in marriage and child births were followed by a "golden age" of early marriage and many children, while starting with the 1950 cohort a massive decline in nuptiality and fertility can be observed. For an intervening period this process was marked by a polarization of family behavior according to the educational level of women (Huinink, 1995; Huinink and Mayer, 1995b). Leaving home has become even earlier for women, but is relatively high for men. Most importantly, as a transition it is more split from other family events, like marriage. The interrelatedness of residential migration with other events in the domains of family or work and the importance of getting one's own home was another topic for the GLHS (Kurz and Blossfeld, 2004; Wagner, 1989). While, as a part of the transitions to adulthood, the family sphere is marked by delays and a certain degree of disintegration (e.g. between cohabitation, marriage and childbearing), changes in the sphere of work have shown remarkable stability over historical time (Brückner and Mayer, 2005).

Life courses in the post-Socialist transformation

Using the retrospective data from the East German Life History Study, we discovered, a worsening of career opportunities and an increasing rigidity of the class structure in the former German Democratic Republic. This might have undermined the legitimacy of the Communist regime and have hastened its downfall (Huinink and Mayer, 1995a; Solga, 1995; Trappe, 1995). In regard to impacts of the East German transformation from Socialism to Social Capitalism and of the unification with West Germany, we found enormous turbulence in work lives (including high exposures to unemployment), but high stability in occupational activity, employment status, as well as family-related social relations. Another remarkable finding was the strong effect transition experiences had in changing control beliefs which are usually assumed to be highly stable and especially salient in times of external crisis (Diewald et al., 2006).

Cross-national comparative studies
The data of the GLHS has been extensively used in cross-national comparative studies, among else in Allmendinger's pathbreaking study on career dynamics in Germany, the US and Norway, in Hillmert's study on education, training and labor market entry in the UK and West Germany (Hillmert, 2001), and the series of studies resulting from the GLOBALIFE Project (e.g., Blossfeld et al., 2005; Blossfeld and Timm, 2003; Drobnic and Blossfeld, 2001; Kurz and Blossfeld, 2004).

6 Data access and documentation

6.1 Documentation

Basic information in German and English, extensive documentation in German as well as lists of publications based on data from the GLHS or other work related to the GLHS project can be accessed at the following Internet sources:

- The home page of the Max Planck Institute for Human Development, Berlin (Germany): http://www.mpib-berlin.mpg.de/forschung/bag/projekte/lebensverlaufsstudie/index.htm
- The homepage of the Center for Research On Inequalities and the Life Course at Yale University, New Haven (USA): http: //www.yale.edu/ciqle/GLHSINDEX.htm

6.2 Data access

The data from the surveys of the German Life History Study are publicly available for scientific research. Requests for the data can be addressed to: Zentralarchiv für Empirische Sozialforschung, Bachemerstrasse 40, D- 50869 Köln (Germany): http://www.gesis.org/za or Center for Research on Inequalities and the Life Course (CIQLE), Yale University, 140 Prospect Street, P.O. Box 208265, New Haven , CT 06250-8265: http://www.yale.edu/ciqle or per email at: ciqle@yale.edu or sarah.gelo@yale.edu

Manuals and extensive documentation in printed form or as CDs are available on request from Redaktion, Max Planck Institute for Human Development, Lentzeallee 94, D – 14195 Berlin (Germany).

Acknowledgements

Parts of the article were written while the author was a visiting professor at the Bavarian State Institute for Family Research at the University of Bamberg (*ifb*) (Germany). Bogdan Mach provided helpful comments.

Glossary

Cohort studies A cohort is a population defined by a specific originating event, like birth in a given year or graduating from high school in a given year. Cohort studies single out one or several cohorts and follow them over time (prospectively or retrospectively). Cohort effects assume that experiences early in life and specific to cohorts or significant parts of it have consequences throughout later life, such as the opportunities restricted by a large cohort size ("baby boom") or the labor market conditions at career entry.

Life course By the term "life course" sociologists denote the sequence of activities or states and events in various life-domains which span from birth to death. The life course is thus seen as the embedding of individual lives into social structures primarily in the form of their partaking in social positions and roles, i.e., with regard to their membership in institutional orders. The sociological study of the life course, therefore, aims at mapping, describing and explaining the synchronic and diachronic distribution of individual persons into social positions across the lifetime. One major aspect of life courses is their internal temporal ordering, i.e., the relative duration times in given states as well as the age distributions at various events or transitions.

Quantitative life histories Life histories can be quantified by distinguishing episodes and states, such as being employed or being married, their historical dating and duration. On this basis events and transitions can be computed as survival distributions, hazard rates and density functions. Quantitative life histories capture the temporal and positional structure of life courses.

Retrospective research design In a retrospective research design a well-defined population, like a cross-section or birth cohort, is sampled and their past history is recorded by relying on the recalling of events and activities. Retrospective measurement can be forward in time, like following an occupational career from the first to the present position, or it can be backward in time from the time-point of the interview, like tracing marriage histories from the present to the first spouse.

References

Allmendinger, J. (1989). Career mobility dynamics: A comparative study of the United States, Norway and Germany, Max Planck Institute for Human Development, Berlin.

Allmendinger, J., Brückner, H., and Brückner, E. (1993). The production of gender disparities over the life course and their effects in old age. Results from the West German Life History Study. In: A. B. Atkinson, M. Rein (Eds.), Age, Work and Social Security, St. Martin's Press, New York, NY.

Becker, R., Schömann, K. (1996). Berufliche Weiterbildung und Einkommensentwicklung, Kölner Zeitschrift für Soziologie and Sozialpsychologie 48(3), 426–461.

Belli, R.F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys, Memory 6(4), 383–406.

Belli, R.F., Shay, W., and Stafford, F. (1999). Computerized event history calendar methods: A demonstration of features, functions, and flexibility. Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Pete Beach, Florida.

Belli, R.F., Shay, W., and Stafford, F. (2001). Event history calendars and question list surveys. A direct comparison of interviewing methods, Public Opinion Quarterly 65(1), 45–74.

Blossfeld, H-P. (1987a). Kohortendifferenzierung und Karriereprozeß – Eine Längsschnittstudie über die Veränderung der Bildungs- und Berufschancen im Lebenslauf, Campus, Frankfurt am Main.

Blossfeld, H-P. (1987b). Zur Repräsentativität der Sfb-3-Lebensverlaufsstudie: ein Vergleich mit Daten aus der amtlichen Statistik, Allgemeines Statistisches Archiv 71, 126–144.

Blossfeld, H-P., Klijzing, E., Mills, M., and Kurz, K. (2005). Globalization, Uncertainty, and Youth in Society, Routledge, New York.

Blossfeld, H-P., and Timm, A. (2003). Who marries whom: Educational systems as marriage markets in modern societies, Kluver, Dordrecht.

Brewer, W.F. (1986). What is autobiographical memory? In: D.C. Rubin (Ed.), Autobiographical Memory, Cambridge University Press, Cambridge, UK.

Brückner, E. (1993). Lebensverläufe und Gesellschaftlicher Wandel Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1919–1921, Max-Planck-Institut für Bildungsforschung, Berlin (materials aus des Bildungsforschung 44).

Brückner, E., and Mayer, K.U. (1998). Collecting life history data: Experiences from the German Life History Study. In: J.Z. Giele, G.H. Elder (Eds.), Methods of Life Course Research: Qualitative and Quantitative Approaches, Sage Publications, Thousand Oaks, CA, pp. 152–181.

Brückner, H. (1995). People don't lie, Surveys do? An analysis of data quality in a retrospective life course study. Materialien aus der Bildungsforschung Nr. 50, Max-Planck-Institut für Bildungsforschung, Berlin.

Brückner, H., and Mayer, K.U. (1995). Lebensverläufe und Gesellschaftlicher Wandel. Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1954–1956 und 1959–1961. Max-Planck-Institut für Bildungsforschung, Berlin (Materialien aus der Bildungsforschung 48).

Brückner, H., and Mayer, K.U. (2005). The destandardization of the life course: What it might mean and if it means anything whether it actually took place, Annual Review of Life Course Research 205(9), 27–54.

Carroll, G.R., and Mayer, K.U. (1986). Job-shift patterns in the Federal Republic of Germany: The effects of social class, industrial sector, and organizational size, American Sociological Review 51, 323–341.

Conway, M.A., and Pleydell-Pearce, C.W. (2000). The construction of autobiographical memories in the self-memory system, Psychological Review 107(2), 261–288.

Diewald, M., Goedicke, A., and Mayer, K.U. (2006). After The Fall of the Wall. Life Courses in the Transformation of East Germany, Stanford University Press, Stamford.

Drobnic, S., and Blossfeld, H-P. (2001). Careers of Couples in Contemporary Societies, Oxford University Press, Oxford.

Featherman, D.L. (1979). Retrospective Longitudinal Research: Methodological Considerations. Journal of Economics and Business 32(2), 152–169.

Goedicke, A., Matthes, B., Lichtwart, B., and Mayer, K.U. (2004). Dokumentationshandbuch – Ostdeutsche Lebensverläufe im Transformationsprozess. Max-Planck-Institut für Bildungsforschung Berlin.

Graaf, P. de. (1987). Intergenerationale klassenmobiliteit in Nederland tussen 1970 en 1986. Mens en Maatschappij 62, 209–221.

Groves, R.M. (1989). Survey Errors and Survey Costs, Wiley, New York.

Henz, U., and Maas, I. (1995). Chancengleichheit durch die Bildungsexpansion. Kölner Zeitschrift für Soziologie und Sozialpsychologie 47(4), 605–633.

Hillmert, S. (2001). Ausbildungssysteme und Arbeitsmarkt: Lebensverläufe in Großbritannien und Deutschland im Kohortenvergleich. Westdeutscher Verlag, Wiesbaden.

Hillmert, S. (2002). Die Edition von Lebensverlaufsdaten – Einzelfallprüfungen, Korrekturentscheidungen und ihre Relevantz, Max-Planck-Institut für Bildungsforschung, Berlin. (Projekt Ausbildungs- und Berufsverläufe der Geburtskohorten 1964 und 1971 in Westdeutchland. Arbeitspapier 20).

Hillmert, S., and Mayer, K.U. (2004). Geboren 1964 und 1971, Verlag für Sozialwissenschaften, Wiesbaden.

Huinink, J. (1995). Warum noch Familie? Zur Attraktivität von Partnerschaft und Elternschaft in unserer, Campus, Gesellschaft. Frankfurt a.M.

Huinink, J., and Mayer, K.U. et al. (1995a). Kollektiv und Eigensinn: Lebensverläufe in der DDR und ach. Akademie-Verlag, Berlin.

Huinink, J., and Mayer, K.U. (1995b). Gender, social inequality, and family formation in West Germany. In: K.O. Mason, A-M. Jensen (Eds.), Gender and Family Change in Industrialized Countries, pp. 168–199. Clarendon Press, Oxford.

Jacob, M. (2004). Mehrfachausbildung in Deutschland: Karriere, Collage, Kompensation? VS - Verlag für Sozialwissenschaften, Wiesbaden.

Jonsson, J.O., and Mills, C. (2002). From Craddle to Grave: Life-course change in modern Sweden. Sociologypress, Durham.

Kurz, K., and Blossfeld, H-P. (2004). Home Ownership and Social Inequality in Comparative Perspective. Stanford University Press, Stanford.

Loftus, E., and Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. Memory & Cognition 11(2), 114–120.

Mach, B.W. (2003). Pokolenie historycznej nadziei i codziennego ryzyka. Spoleczne losy osiemnastolatkow zroku 1989 [Generation of Historic Hope and Everyday Risk. Social Trajectories of Eighteen-year Olds from the Year 1989]. Institute of Political Studies of the Polish Academy of Sciences, Warsaw.

Mayer, K.U. (1990). Lebensverläufe und sozialer Wandel: Anmerkungen zu einem Forschungsprogramm. In K.U. Mayer (Ed.), Lebensverläufe und sozialer Wandel. Kölner Zeitschrift für Soziologie und Sozialpsychologie: Sonderhefte 31, 7–21. Westdeutscher Verlag, Opladen.

Mayer, K.U. (2000). Promises fulfilled?: A review of 20 years of life course research, Archives Européennes de Sociologie 41, 259–282.

Mayer, K.U. (2001). The paradox of global social change and national path dependencies: life course patterns in advanced societies. In: A. Woodward & M. Kohli (Eds.), Inclusions and Exclusions in European Societies, Routledge, London, pp. 89–110.

Mayer, K.U. (2004). Whose lives? How history, societies, and institutions define and shape life courses, Research in Human Development 1(3), 161–187.

Mayer, K.U., and Aisenbrey, S. (forthcoming). Variations on a major theme – Trends in social mobility in (West-) Germany for cohorts born between 1919 and 1971. In: M. Gangl und S. Scherer (Hrsg.), Festschrift für Walter Müller, Campus Verlag, Frankfurt.

Mayer, K.U., and Baltes, P.B. (Eds.). (1996). Die Berliner Altersstudie, Akademie Verlag, Berlin.

Mayer, K.U., and Brückner, E. (1989). Lebensverläufe und Wohlfahrtsentwicklung: Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929–1931, 1939–1941, 1949–1951. Max-Planck-Institut fur Bildungsforschung, Berlin (Materialien aus der Bildungsforschung 35).

Mayer, K.U., and Carroll, G.R. (1987). Jobs and classes: structural constraints on career mobility. European Sociological Review 3, 14–38.

Mayer, K.U., and Huinink, J. (1990). Age, period, and cohort in the study of the life course: A comparison of classical A-P-C-analysis with event history analysis or farewell to Lexis?. In: D. Magnusson, L.R. Bergman (Eds.), Data Quality in Longitudinal Research, Cambridge University Press, Cambridge.

Mayer, K.U., and Müller, W. (1986). The state and the structure of the life course. In: A.B. Sorensen, F.E. Weinert, L.R. Sherrod (Eds.), Human Development and the Life Course: Multidisciplinary Perspectives. Erlbaum, Hillsdale, NJ, pp. 217–245.

Mayer, K.U., and Solga, H. (1994). Mobilität und Legitimität: zum Vergleich der Chancenstrukturen in der alten DDR und der alten BRD oder: Haben Mobilitätschancen zu Stabilität und Zusammenbruch der DDR beigetragen?, Kölner Zeitschrift für Soziologie und Sozialpsychologie 46, 193–208.

Pollmann-Schult, M., and Mayer, K.U. (2004). Returns to skills: Vocational training in Germany 1935–2000, Yale Journal of Sociology 4, 73–98.

Reimer, M. (2001). Die Zuverlässigkeit des autobiographischen Gedächtnisses und die Validität retrospektiv erhobener Lebensverlaufsdaten. Kognitive und erhebungspragmatische Aspekte. Max-Planck-Institut für Bildungsforschung, Materialien aus der Bildungsforschung Nr. 71, Berlin.

Reimer, M. (2005a). Autobiografisches Gedächtnis und retrospektive Datenerhebung. Die Rekonstruktion und Validität von Lebensverläufen, Max-Planck-Institut für Bildungsforschung, Studien und Berichte Nr. 70, Berlin.

Reimer, M. (2005b). Collecting Event History Data About Work Careers Retrospectively: Mistakes that Occur and Ways to Prevent Them. Max Planck Institute for Human Development, Berlin.

Reimer, M., Matthes, B., and Pape, S. (forthcoming). Collecting Event Histories with True Tales: Techniques to Improve Autobiographical Recall Problems in Standardized Interviews, Quality and Quantity.

Rubin, D.C. (Ed.) (1986). Autobiographical Memory. Cambridge University Press, Cambridge.

Rubin, D.C. (Ed.) (1996). Remembering Our Past: Studies in Autobiographical Memory. Cambridge University Press, Cambridge.

Schwarz, N., and Sudman, S. (Eds.) (1996). Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, Jossey-Bass, San Francisco.

Solga, H. (1995). Auf dem Weg in eine klassenlose Gesellschaft?: Klassenlagen und Mobilität zwischen Generationen in der DDR. Akademie-Verlag, Berlin.

Solga, H. (1996). Lebensverläufe und Historischer Wandel in der Ehemaligen DDR. ZA-Information 38, 28–38.

Solga, H. (2001). Longitudinal surveys and the study of occupational mobility: Panel and retrospective design in comparison, Quality and Quantity 35(3), 291–309.

Solga, H. (2005). Ohne Abschluss in die Bildungsgesellschaft: die Erwerbschancen gering qualifizierter Personen in soziologischer und ökonomischer Perspektive, Barbara Budrich, Opladen.

Sudman, S., Bradburn, N.M. (1996). Thinking About Answers: The Application of Cognitive Processes to Survey Methodology, Jossey-Bass Publishers, San Francisco.

Trappe, H. (1995). Emanzipation oder Zwang?: Frauen in der DDR zwischen Beruf, Familie und Sozialpolitik, Akademie-Verlag, Berlin.

Wagner, M. (1989). Räumliche Mobilität im Lebensverlauf. Enke, Stuttgart.

Wagner, M. (1996) Lebensverläufe und Gesellschaftlicher Wandel. Die Westdeutschen Teilstudies. ZA-Information 38, 20–27.

Wehner, S. (2002). Exploring Trends and Patterns of Nonresponse: Evidence from the German Life History Study. Essex Summer School in Social Science Data Analysis and Collection, University of Essex, Colchester.

Part II

Measurement Issues in Longitudinal Research

This page intentionally left blank

Chapter 7

Respondent recall

Jennifer K. Grotpeter

1 Introduction

How did we become the people we are today? An interest in human development is as old as humanity itself, and the scientific study of human development began with the publication of Charles Darwin's *Origin of the Species* in 1859. Since then, the evidence used to support developmental theories has evolved from field observation notes to complex longitudinal studies following infants and children into adolescence and adulthood. As the life course perspective of human development has progressed since the 1960s, researchers have become increasingly aware that our lives are individually shaped by the timing and sequencing of our life events (e.g., Elder and O'Rand, 1995).

In order to conduct scientific inquiry assessing those life events, clearly researchers must be able to measure study participants' characteristics, attitudes, and behavior throughout the lifespan. Because it is impossible to follow study participants from birth to adulthood, recording their every thought and behavior, along with everything that happens to them, researchers must rely on asking participants about their thoughts, behaviors, and the events that happened to them over some period of time. Whether this time period is the past 24 hours, the past year, or the past 20 years, all developmental and life course research is dependent upon the ability of its subjects to accurately recall what they have previously thought and done. This chapter will explore issues in respondent recall in longitudinal survey research, focusing particularly on issues of reliability and validity in short-term and long-term retrospective recall.

2 Respondent recall – issues of memory

Research on memory indicates that we are unlikely to completely forget things that we have directly experienced, though we may experience problems with recalling these events accurately (Fowler, 1998). Megan Beckett and colleagues, in 2001, reviewed the literature on error in respondent recall, distinguishing between retrieval of autobiographical memories, recall strategies and recall error, and respondent characteristics associated with recall error. The following discussion of recall error is based on the framework used by Beckett, et al. (2001), and is supplemented by other reviews.

2.1 Retrieval of autobiographical memories

In their review, Beckett and colleagues (2001) found that researchers have identified four steps in the reporting process, which is the same for retrieval of past events, current status reports, and attitudinal questions: (1) the respondent interprets the question; (2) the respondent retrieves the answer to the question or information needed to construct an answer; (3) the respondent formulates an answer based on the recalled information; and (4) the respondent edits the answer and decides whether and how to respond. Most research on recall error focuses on the second part of the process, memory retrieval. The most obvious recall problem, "forgetting," occurs when events were never processed, encoded, or stored into long-term memory in the first place. Alternatively, they argue that the major event may be firmly encoded into memory, but the finer details may not be easily recalled and are gradually lost over time.

The second recall issue that Beckett and colleagues (2001) discussed is that the event may not be recalled due to retrieval error, which means that the event is in the memory, but because too much time has passed since it was "rehearsed," (i.e., it needed to be remembered or thought or talked about), it was too difficult to access the memory. Birthdates and wedding anniversaries are generally rehearsed regularly, whereas the date one had coffee with a friend is not. The mental processes involved in rehearsal are thought to strengthen the pathway to the memory, which in turn increases the ease of later recall. Third, the review found that recall problems may be due to the inaccurate reconstruction of a memory, due to other similar events "overwriting" the initial memory.

In reviews, Fowler (1998), Charles Pierret (2001), and Beckett and colleagues (2001) assert that another threat to accurate recall is the length of the recall period (i.e., the time between the event and the time it is reported). A longer recall period increases the quantity of information the respondent must retrieve from memory, making it more difficult to do so, and the high quantity of information in turn increases the likelihood that the respondent will be unable to distinguish between various events, confusing details between two events or forgetting one event entirely (Pierret, 2001). In their First and Second Malaysian Family Life Surveys (MFLS-1 and MFLS-2), Beckett and colleagues (2001) studied ever-married women of childbearing age and their husbands. One sample was drawn for the first study and then 72% of the respondents in that sample were interviewed at the second time, at which time a second sample was newly drawn and the respondents were given retrospective interviews on topics such as breastfeeding their children. Results indicated that data quality did deteriorate with the length of the recall period. They reported that some events were forgotten and details about other events were "blurred," resulting in incomplete or inexact reports.

The time of recall is not simply a linear issue, however. Fowler (1998) notes that the greater the impact or current salience of the event, the more likely it is to be recalled. Similarly, in their literature review, Beckett and colleagues (2001) found that the rate of decline in recall ability over time varied by type of event. That is, recall of some types of events that were more infrequent and salient, such as annual physicals and reporting of robberies, did not decay over time, but that reports of assaults, burglaries, and larcenies did suffer from memory decay over time. They note in particular that in these instances, events in the distant past are prone to suffer from "heaping," meaning that if respondents could not recall an actual value, they provided a "prototypical" response nearest the actual value, which results in an artificially high number for the given response period.

Though salience may seem to be related to the rehearsal issue discussed above, it is distinct in that salience depends upon the relationship of the event to the respondent's lifestyle or self-identity. For example, Beckett cites her own 2000 study of older Taiwanese, in which the most important predictor that a health condition would be reported at both time periods studied was if that condition strongly affected their daily lives, such that it affected walking or bathing (Beckett, Weinstein, and Yu-Hsuan, 2000). Beckett, et al. (2001) also found using their Malaysian data that among men and women, moves that were associated with highly salient life events (e.g., new jobs for men or marriage for women) were more likely to be consistently reported 12 years later.

A third recall problem is referred to as telescoping (Beckett, et al., 2001; Pierret, 2001). Telescoping is said to occur when more events are recalled as having occurred in the most recent period and fewer events in the more distant past. Three factors Beckett and colleagues (2001) report as contributing to telescoping are that (1) normal forgetting is greater for events that occurred in the more distant past; (2) errors in dating events that may otherwise be random or unbiased increase over time; and (3) events that occurred outside the reporting period may intrude upon the recall of events in the reporting period (i.e., respondents may be asked about the number of times they argued with their spouse in the past six months, but they may include arguments that occurred between six and twelve months ago). The second and third factors discussed above may then lead to overestimation of the number of events in a recent reporting period. In their 2001 Malaysian study, Beckett confirmed that there was a slight telescoping of events in the second survey compared with reports of the same events in the first survey 12 years earlier. A similar issue discussed in the review is referred to as the "availability heuristic," or "availability principle." This term indicates that if an event is easily recalled, the respondent may believe it happened with more frequency than if the event were more difficult to recall.

Finally, when a respondent is unable to recall events clearly, several strategies are regularly used to aid in reconstructing the events, but some of these strategies themselves may lead to reporting bias (Beckett, et al., 2001). One common strategy is to probe for individual instances of the event, but this may lead to "heaping." Another strategy is to use information about current attitudes or behavior to infer past attitudes or behavior, but this results in "stability bias," which artificially diminishes potential changes in attitudes and behavior over time.

Fowler (1998) states that the more consistent an event was with the way the respondent thinks about things, the more likely it is to be recalled. While retrospective reports were generally found to be strongly influenced by current behaviors, Beckett and colleagues (2001) also found that in some instances, retrospective reports may be more accurate than concurrent reports when the issues are particularly sensitive, such as heavy drinking—respondents were more willing to acknowledge such events as having occurred in the distant past rather than as a current embarrassing behavior. In the Malaysian study, Beckett, et al. (2001) found that even reports about less socially sensitive behaviors, such as breastfeeding, were responsive to social norms similarly to more sensitive topics (e.g., abortion or illegal drug use).

Finally, Beckett, et al. (2001) discuss that certain sociodemographic subgroups, such as better-educated women, may be better able to recall past use of health care, and that this was more true for events in the more distant past than the more recent past. In their Malaysian studies, Beckett, et al. (2001) found that a correlate of data quality, such as education, in retrospective reports on one topic (e.g., details about a birth) is likely to be predictive of quality issues related to another topic (e.g., migration histories). It is also possible in a longitudinal survey that respondents become aware over the years that the more questions to which they respond affirmatively, the more follow-up questions they will be asked. They may elect to reduce their effort and not report certain events in order to be spared being asked additional questions. Further discussion of this phenomenon of *panel conditioning* may be found in the next chapter (Cantor, Chapter 8 in this volume).

Over thirty-five years ago, Marian Radke-Yarrow and colleagues (Radke-Yarrow, Campbell, and Burton, 1970) compared contemporaneous and retrospective accounts of childhood behavior and experiences, and the results revealed biases and distortions that occur with research based on retrospective recall. In 1998, Sir Michael Rutter and colleagues (Rutter, Maughan, Pickles, and Simonoff, 1998) conducted a follow-up literature review of studies that followed Radke-Yarrow's work. In their review of literature on memory, Rutter and colleagues (1989) discuss the extent to which, assuming that accurate memories are available in the mind for recall, the reconstruction of memory introduces systematic bias. In particular, it is argued that people use "implicit theories" (personal meaning about specific phenomena) to reconstruct their personal histories, and there is strong evidence that reports of past behavior may be distorted to better reflect current feelings and attitudes. However, evidence is also presented that with skillful interviewing some of these biases may be overcome, and that distortion of memory is not necessarily inevitable.

3 Respondent recall – issues in longitudinal research design

Information gathered at multiple points in time is crucial for any research that aims to predict future behavior based on past behavior and beliefs. Thus, multiple measurement points are necessities for studies of lifespan development studies that are designed to assess developmental pathways from childhood to adulthood, and studies that require lifetime estimates of conditions, such as psychopathology (Rutter, Maughan, Pickles, and Simonoff, 1998). There are two primary ways in which researchers can assemble longitudinal datasets – through the use of prospective (longitudinal) panel designs, and through the use of cross-sectional retrospective designs. Though both designs require respondents to recall back some amount of time in order to generate a response, they each necessitate substantially different lengths of recall time.

3.1 Prospective panel designs

Many of the predictive antecedents of later behaviors or events in people's lives are best measured at the time they occur. Thus, a longitudinal design that follows a sample over time, concurrently measuring beliefs, events, and behavior, is ideal for assessment using a life course perspective. In prospective longitudinal, or panel, research, a subject pool is identified and baseline interview data is gathered. This same subject pool is tracked over time and is then re-contacted and interviewed at specified intervals. At the conclusion of the study, optimally, data will be available for every subject at each point in time, so that potential changes can be examined over time, with relatively minimal reliance on the extreme long-term memories of the participants.

3.2 Cross-sectional retrospective designs

The second way that researchers can assemble a type of longitudinal dataset is by using a cross-sectional retrospective design, which is designed to significantly shorten the amount of time required to gather longitudinal data, as well as to conserve financial and human resources. As the name implies, studies using this method combine a cross-sectional design with retrospective recall in order to produce longitudinal data. Thus, respondents are surveyed at one point in time and are asked to report about current and past events, behavior, and beliefs. For example, adults may be asked to think back to childhood and to report on their peer relationships or their mother's parenting style, or they may be asked to report the age at which they first drank alcohol, smoked cigarettes, or engaged in sexual behavior. Respondents may be asked to recall back for a length of time up to several decades (e.g., the

Health and Retirement Study). As a result, in one wave of data collection, a researcher gains access to data that refers to multiple points in the respondents' lives, potentially over a very long time span, that would otherwise take many years and a great deal of financial and human resources to collect.

It is important to note that researchers must consider respondent recall issues even in prospective research because these studies gather data on time periods ranging from the past few months (e.g., Survey of Income and Program Participation), to the past year (the Panel Study of Income Dynamics, the National Youth Survey), to several years (e.g., Wisconsin Longitudinal Study). The important question is, what is the length of time over which events can be recalled with reasonable reliability? Scott and Alwin (1998) argue that the optimal design for gathering life history data is one that combines the best features of prospective and retrospective measurement. As a result, understanding the issues in respondent recall is crucial for all longitudinal researchers, including those who conduct prospective research, and for anyone desiring to comprehend the ramifications of the results of that research.

3.3 Benefits and limitations – prospective designs

Jacqueline Scott and Duane Alwin (1998) state the following benefits of prospective designs: (1) they allow for data to be collected concurrently with the events in question; (2) they allow for the continuous measurement of events and changes that would be too burdensome in retrospective studies; (3) it is much easier for researchers to impose theoretically driven definitions of life events that are ambiguous in practice, such as "leaving home"; (4) they provide an opportunity to collect prospectively oriented data concerning individual aspirations and expectations and to compare these to actual outcomes at later points in time; and (5) they provide data on the interaction of individual life trajectories within the household or family context, whereas retrospective designs would be limited to only those households that survive.

While the prospective panel research design is the ideal design for measuring change over time, there are some significant limitations. First, as Scott and Alwin (1998) argue, statistical problems accrue because of subject attrition or non-response, and this attrition may not be random. Second, it is possible that interviewing the same people over time affects the essence and quality of respondents' answers. This concern with panel conditioning is somewhat similar to the "Hawthorne effect" whereby the act of making people subjects on a social experiment affects their subsequent behavior and makes them less "typical" (Scott and Alwin, 1998).

Additionally, there are several practical limitations to prospective panel studies. First, these designs require substantial investments of financial and human resources, and they require a great deal of time (i.e., from years to several decades). Second, if the study continues a decade or more, it is likely that new constructs will be identified and old constructs will change within the research field, and it is impossible to go back in time to the first year of the study and change or add questions to earlier surveys. Similarly, because it is the ideal in a prospective panel study for individual questions to remain constant across the waves of the study so that changes in responses to those questions can be measured over time, it is rare that a researcher would sacrifice this consistency to change the wording of a question, even if the reason were to improve a poorly worded question or to follow new developments in research. Despite these limitations, prospective designs remain the most reliable and valid way to amass longitudinal data that can be used to assess temporality and causal relationships.

3.4 Benefits and limitations – retrospective designs

Retrospective research designs have two practical benefits over prospective designs. First, financially, it is far less expensive to survey adult respondents at one time point to answer questions about their life experiences over the past twenty years than it is to identify, track, and survey the same people from childhood through adulthood to ask them the same questions every year. Second, temporally, obtaining longitudinal results using a one time retrospective interview takes only as much time as it takes to conduct the interview and analyze the data, compared with the years (and even decades) required before the data reward of a prospective longitudinal study is available.

The most obvious drawback to retrospective research designs is that this method relies heavily on the accuracy of respondent recall, which is affected both by memory recall issues and the way in which they filter their memories based on current beliefs. Additionally, these recall issues may negatively impact the recall of some types of events and behaviors more than others, and so it is important for the researcher considering this type of research design to be aware of the type of data they intend to collect and the degree to which it is susceptible to recall problems. Another drawback of retrospective designs is that they can result in selection biases because only survivors can be interviewed (Scott and Alwin, 1998). Finally, Scott and Alwin (1998) argue that recent experiences and events may bias the recollections people make about their earlier experiences, making inferences about trends or causation somewhat circular.

Thus it is clear that with unlimited financial, temporal, and human resources, it is generally better to collect information from individuals prospectively, i.e., "to gather data on people's lives as they are living them" (Scott and Alwin, 1998). Despite this, because of limited resources, much criminological research makes use of retrospective recall in cross-sectional research designs, and it is important to understand the degree to which these data can provide accurate measures of the topics in question. "The challenge to those designing the optimal longitudinal measurement design, given considerations of cost and data quality, is to find the acceptable limits for gathering data retrospectively" (Scott and Alwin, 1998).

4 Research evidence

The following section contains brief descriptions of a variety of studies that used prospective or a combination of prospective and retrospective research designs, and raise or address issues in recall.

To address the debate about the utility of longitudinal compared to cross-sectional research designs, Scott Menard and Delbert Elliott (1990) examined empirical evidence on the extent to which longitudinal and cross-sectional data could be used interchangeably without affecting substantive conclusions. Data were taken from the National Youth Survey (NYS), a nationally representative sample of adolescents who were aged 11 to 17 when the sample was drawn in 1976. The NYS used annual assessments for the first five years of the study. The next two assessments occurred three years after the last annual assessment, and then again three years after that. Though the focus for all seven waves of the study was on the previous year, at wave six some retrospective data were also collected for the other two intervening years, which constituted two- and three-year recall periods. This allowed analysis of delinquency and drug use variables for a single wave (i.e., cross-sectional data) and also age-specific estimates for each birth cohort over multiple waves (i.e., longitudinal data collection). Results based on two different analyses comparing prospective data with short-term retrospective data indicated that in comparison with the prospective longitudinal panel, the retrospective cross-sectional data collection that involved extended recall periods significantly underestimated the prevalence of delinquency and drug use. This underestimation was more serious for general delinquency and Index offenses than it was for marijuana use or polydrug use.

Over longer periods, the problem of comparing long-term retrospective data with longitudinal data appeared to be even more severe. In the seventh wave, the respondents were asked about whether they had ever engaged in a series of offenses, including the Index offense scale, and if they had, at what age had they first committed the offenses. Of those who reported ever having committed an offense on either the prospective or the retrospective questions, about 10% failed to do so on the prospective surveys, and more than half failed to do so on the retrospective surveys. In two-thirds of the cases in which an offense was reported on either of the two measures, the two measures disagreed about whether or not an offense had ever been committed. In an additional 11% of these cases for which an offense was reported on either of the two measures, the two measures disagreed on the age at which the respondent first committed the offense. Overall, it was only in one-fourth of the cases that prospective and retrospective accounts agreed that an offense was committed at a particular age. The data used in this study demonstrated clearly that longitudinal data collected using cross-sectional methods with extended recall periods may produce very different results from longitudinal panel data collected prospectively.

Kevin Weinfurt and Patricia Bush (1996) observe that there is an assumption in longitudinal studies of childhood substance use that repeated measurement of a particular variable for a particular child will be “reliable and noncontradictory” across sampling waves. However, they note that such assumptions are rarely examined, except perhaps as internal consistency in cross-sectional designs. The authors propose that, alternatively, external consistency, which is the change in response patterns over time, should be examined. Using a framework presented by Bailey, Flewelling, and Rachal (1992), they examined logical errors (errors in which the respondent indicates engagement in some behavior at an earlier time point and no ever-engagement in that behavior at a later time point) and estimation errors (errors in which a respondent indicates less ever-engagement at a later time period than they indicated at an earlier time period) committed by elementary and junior high school students who were administered drug use surveys every year for four years. Students were asked to indicate the number of times they had more than one puff of a tobacco cigarette, more than a puff of marijuana, or more than a sip of alcohol. It was unclear what the time bounds were for these responses (e.g., in the last month, six months, or past year).

Logical errors tended to decrease for all substances as the students became older, while the percentage of estimation errors remained stable (alcohol and cigarettes) or increased slightly (marijuana). As a result, the ratio of logical to estimation errors decreased across the survey years. Both logical and estimation errors were found for every substance in every lag of the study, but whereas estimation errors remained relatively stable over time, logical errors decreased for all substances as students got older. Interestingly, the logical errors for marijuana use were quite high in the first two years of the study, indicating that over half of students who reported marijuana use while in fourth or fifth grade reported no use the following year, which would indicate, if it were the case that student reports become more accurate as they get older, that nearly half of marijuana use reported in fourth and fifth grades was a false positive report. Apparently, counterintuitively, shorter time lag between responses resulted in worse errors than longer time lag. However, the authors note that any estimation of error that involved Time 1, or the youngest respondents, resulted in the greatest errors, and so it is confounded with age of the subjects.

Deborah Freedman and colleagues (Freedman, Thornton, Camburn, Alwin, and Young-DeMarco, 1988) used a life history calendar (LHC, detailed later in this chapter) in order to aid respondent recall. These calendars were used in a validation study with 900 23-year-olds who were asked to report about the nine years since their fifteenth birthdays. Data from initial interviews from one particular month in 1980 were compared with retrospective recall from 1985 of that same month in 1980. The authors reported that of the 900 LHCs, only four calendars had months with no data, and the 1985 retrospective data corresponded highly with the 1980 interview data and interviewers evaluated the procedure favorably. Overall, those activities with a relatively high level of volatility, such as employment activity, were least consistent between the two reporting periods, and thus the authors note that researchers should be aware the highly variable events will likely be measured less accurately than more stable events. (On a related note, Henry, Moffitt, Caspi, Langley, and Silva (1994) found that retrospective measures of psychological or attitudinal variables were less consistent with prospective data than were measurements of more objective characteristics such as changes in residence or police contacts.) A cautionary note, however, is sounded by Taris (2000), who concluded after a review of research on the LHC that the LHC improved information only some of the time and for some but not other variables.

In order to examine the quality of information that could be gathered on lifecycle events in one retrospective survey, H. Elizabeth Peters (1988) conducted a study that compared data from a retrospective marital history with those derived for the same individuals from panel information. Data were taken from the Young Women's cohort of the National Longitudinal Surveys of Labor Market Experience (NLS). In 1978, respondents were asked about the dates of past marital events (i.e., marriage, divorce, remarriage) and husbands' characteristics (for verification) for each marriage, and these retrospective marital histories were updated five years later. Meanwhile, approximately annual panel information was also available with similar survey questions (i.e., current marital status). The authors found that this panel information sometimes yielded results that were difficult to interpret, because it was not necessarily clear if more than one event of short duration had occurred in one year (e.g., the subject was married in one year, and in the next year they divorced and remarried, but in both assessments they were coded as "married"). Thus, though the retrospective responses were more prone to memory error, these data tended to be more complete.

Results indicated that when a marital event was reported in both the retrospective and panel sources, there was substantial agreement about the dates of the event, and when there were errors, they appeared to be related to difficulty of recall in the retrospective histories. Specifically, there was very little discrepancy between the reports of age at first marriage between the retrospective and panel sources; however, subsequent divorce and remarriage analysis, while qualitatively similar, yielded less precise parameter estimates. When the time between the date of the event reported in the retrospective survey and the date of the actual survey was greater, there was a substantially greater likelihood of inconsistent information. The authors argue that those characteristics that vary over time, such as labor force participation, experience, and earnings, are usually not reported in retrospective surveys because information about the timing and amounts would be much less reliable than marital histories.

In 2001, Charles Pierret published an analysis comparing annual and biennial responses to the National Longitudinal Survey of Youth 1979 (NLSY79). The NLSY79 made this change in 1994, and at this time asked respondents to report on 1992 despite the fact that they had already completed interviews on 1992 in 1993. The NLSY79 is a panel study of men and women who were aged 14 to 21 in 1978, assessing a variety of topics including schooling, employment, health, marriage, fertility, income, program participation, and crime and illicit activities. By increasing the reporting period to two years, the researchers anticipated several threats to the accuracy of the reports. There was concern that with a longer reporting period, it would be more likely that recall would decay, and that there would be more events to recall. Finally, there was concern that similar or related events would occur during the reference period, causing difficulties in distinguishing between the events, leading to omission of an event or confusion between multiple events.

On their face results indicated that the switch from annual to biennial reporting would have only moderate impacts on event history data. Overall, receiving food stamps and AFDC (Aid to Families with Dependent Children, i.e., welfare) was somewhat under-reported, and benefits were moderately over-reported in the longer reporting periods. However, at the individual level, discrepancies were cause for greater concern. Over two-thirds of AFDC and food stamp recipients reported a different assistance receiving history when asked after one versus two years' time. Additionally, one out of three new employers were not reported, and new dates were given for fully half of all dates reported.

5 Research techniques for improving respondent recall

It seems evident that the best way to optimize the accuracy of respondent recall is to minimize the length of time between the events to be recalled and the time of the interview. The maximum recall time that can be expected to yield reliable, valid results varies by the level of detail that must be recalled. For example, researchers who assess dietary intake have found that even a 24-hour recall period can result in reporting error due to inaccurate recall of food consumed (Fowler, 1998). Alternatively, given that even the most highly regarded prospective longitudinal studies of substance abuse and delinquency involve annual assessments, it is important to maximize reporting accuracy by using techniques to improve recall. The following are techniques which have been suggested to improve respondent recall:

1. Ask long, rather than short questions. Fowler (1998) notes that this does not mean asking convoluted questions, but instead including introductory material that will prepare the respondent for the question. In addition to preparing the respondent for the topic area of the question to follow, this increases the time that the respondents have to search their memories.
2. First ask a summary question (Beckett, et al., 2001). Details on sensitive topics may be better recalled if the respondent is first asked one or more summary questions. For example, Beckett reports that in the MFLS-2, rather than immediately asking respondents specific questions about all pregnancies they have ever had, respondents were first asked about the total number of pregnancies of living and deceased children they have had. Similarly, in the National Youth Survey, respondents are first asked about the total number of children they have before they are asked any specific details about their children.
3. Ask multiple-related questions, which will improve the probability that an event will be recalled and reported (Fowler, 1998; Rutter, et al., 1998). This is effective in part

because it stimulates associations with what the respondent is supposed to report.

4. Train interviewers to ask alternative questions if a respondent is unable to provide a precise response (Beckett, et al., 2001; Freedman, et al., 1988). Data quality for events in the more distant past may be improved by asking alternative questions when necessary. For example, if a respondent cannot recall an exact date, a follow-up question about the respondent's age at the time of the event, or asking for an approximate answer if a specific answer cannot be recalled, may yield a usable answer.
5. Similarly, train interviewers to prompt with consistent follow-up questions (Beckett, et al., 2001; Rutter, et al., 1998; Freedman, et al., 1988). Beckett, et al. (2001) argue that well-trained interviewers may be able to elicit information about underreported events. Specifically, they note that MFLS interviewers were trained to probe further about "miscarriage, induced abortion, missed episodes of contraceptive use, or spousal separation if the period between births was greater than four years." Additionally, Rutter, et al. (1998) note that this facilitates separation of attitudes from behavior, relating an anecdote from a prior study in which a woman replied that her husband "did nothing" around the house, but when probed further with specific questions about activities during the previous week, it became apparent that he was very much involved around the house, which she then acknowledged with surprise. The initial answer reflected her negative attitude toward her partner while the probes yielded a more factual account of his involvement in household tasks. A drawback to this technique is that interviewers must probe consistently, otherwise some respondents will essentially receive different interviews than the rest of the respondents.
6. Use a calendar, or possibly a life history calendar (LHC; Freedman, et al., 1988; Beckett, et al., 2001; Fowler, 1998; Rutter, et al., 1998; Scott and Alwin, 1998). Recall may be improved by providing respondents with a calendar on which major national and socio-cultural events are pre-printed. This may improve the internal consistency and sequencing of the respondent's answers because they stimulate recall activities that help them place events in time, and because they help to generate boundaries for reporting periods. Because LHCs are physical, visual aids, they can help respondents to relate visually and mentally to the timing of different events. LHCs can be bounded in any time unit needed – day, week, month, or year – and they may make use of a variety of substantive domains as defined by the need of the study, including geographical residence, marital and cohabitation statuses and transitions, fertility, school enrollment, and employment. Variables may be categorical, ordinal, or interval. Interviewers must be trained in assisting the respondent with completing an LHC, which may constitute the entire interview, or may be integrated into a total interview format. If the LHC is integrated into a larger interview, Freedman, et al. (1988) note that the LHC would typically comprise the first part of the interview. If, however, one elects to intersperse calendar questions with non-calendar questions, this may potentially aid recall and time sequencing in the traditional interview section, but it may also confuse the respondent. A potential practical drawback of the calendar method is that as the complexity of the calendar increases, so does the complexity of coding responses, resulting in a potentially tedious (and expensive) coding process. Scott and Alwin (1998) took the life history calendar a step further, and argued that life histories should also assess such constructs as past levels of psychological

well-being. Thus, in order to represent this broader view of life history data, they suggest that the measurement should include variables that fall into three categories: (1) event histories (i.e., the collection of past events, including their timing, duration, and sequencing, and also present statuses, and future expectations); (2) the accumulation of experiences (i.e., where people are at a particular point in time, including the accumulation of experiences that result from people's event histories and experiences resulting indirectly from life events, such as amount of schooling, and expectations for future experiences); and (3) the evaluation or interpretation of experiences (in either the present, the past, or as anticipated in the future).

7. Provide benchmarking (from Beckett, et al., 2001; Fowler, 1998; Pierret, 2001). Respondent recall may be improved by reminding them of their situation at the time of the previous wave and bringing them forward from that point. Several large, national surveys, including the Panel Survey on Income Dynamics and the Health and Retirement Study, take advantage of computer technology to integrate benchmarking into subsequent interviews. Beckett, et al. (2001) note a number of caveats for this method: (1) if the question involves a concept about which more is known at the later time period than the earlier time period, benchmarking may confuse the respondent more than help them; (2) benchmarking requires financial resources, as the interviewer must have access to responses to earlier waves when conducting the interview; (3) benchmarking works best when it works forward in time from the previous wave, while retrospective studies often work backward from the current interview; and, most importantly, (4) if a respondent makes a reporting error at one time period, that error will be carried forward into future waves and may not be corrected because the earlier reports become the "correct" reports. A potential drawback of this method is that it can create a "stability bias," creating more consistency across time than was actually there.
8. Vary recall period with saliency (Beckett, et al., 2001). Because more salient life events are better recalled, the recall period can be restricted for common, nonsalient events, whereas more salient events can be collected for longer recall periods.
9. If the goal is to gain a broad understanding of the respondent's life course, gather data on a range of potentially relevant experiences of a major kind using a mixture of more open questioning about important domains and systematic questioning of details (Rutter, et al., 1998).

6 Conclusion

Survey data on cognitive states can be collected in a design that is truly prospective, but data on experiences or behavior must necessarily be, at least to some extent, retrospective. All other things being equal, longer recall periods appear to generate less accurate results than shorter recall periods. To the extent that the use of longer recall periods in longer term retrospective studies can be justified, the justification is based not on improved or equal accuracy, but on such considerations as cost, and the time between initial data collection and availability of the data for analysis and dissemination of the results. This does not mean that we should abandon longer term retrospective studies because poor quality data are inevitable in such studies. As described here and by Mayer in Chapter 6 of this volume, it is possible in studies with longer recall periods to take measures to improve respondent recall over longer spans of time. What is important is to recognize the inherent problems in recall for any length of time, the magnification of the problem of accuracy when recall periods become longer, and the need to match appropriate techniques to enhance recall with the length of recall required in a specific study. For purposes of maximizing accuracy of recall in longitudinal survey research in general, however, it remains the case that minimizing the time between the events or behaviors of interest and the questions about those events or behaviors, when feasible, is preferable to the use of longer term retrospective data.

Glossary

Cross-sectional retrospective research design This is research that uses a cross-sectional design, but by the use of retrospective recall methods, gathers longitudinal data. These data are designed to represent attitudes, behaviors, and events in the respondents' lives across time, despite the fact they are collected at a single point in time.

Heaping A phenomenon that occurs in memory recall when respondents cannot recall a specific value, so instead they provided a "prototypical" response near the actual value. As a result, certain dates, ages, durations, or frequencies may be over-represented.

Life history calendar (LHC) A life history calendar is an interview technique that assists respondents in supplying accurate information on events that occurred in the past. LHCs include references to national holidays and other known events, and may be bounded by the time unit most appropriate to the study (such as a week, a month, or a year). LHCs can help respondents to relate visually and mentally to the timing of different events, which should improve their ability to recall events and to place those events in their proper time sequence.

Panel study A longitudinal study in which a panel of individuals is interviewed at intervals over a period of time.

Prevalence A type of data that reflects whether or not an event has happened, or whether or not a respondent has engaged in a behavior. It does not address how often the event or behavior may have occurred.

Prospective reporting Prospective means looking to the future. In a prospective panel study, measurement occurs with each wave of the study, moving forward (even though the data collected are about the present or relatively recent past).

Reliability Reliability is the extent to which a construct is consistently measured without random measurement error.

Retrospective recall Thinking about, remembering, and reporting events that happened in the past.

Telescoping A phenomenon that occurs in memory recall that refers to the allocation of events, characteristics, or behaviors to a more recent time period than the one in which it actually occurred.

Validity This is concerned with assessing whether what actually is being measured is related to some external reality. That is, are we measuring what we think we are measuring?

References

Bailey, S. L., Flewelling, R. L., and Rachal, J. V. (1992). The characterization of inconsistencies in self-reports of alcohol and mariquana use in a longitudinal study of adolescents. *Journal of Studies of Alcohol*, *53*, 636–647.

Beckett, M., DaVanzo, J., Sastry, N., Panis, C., and Peterson, C. (2001). The quality of retrospective data: An examination of long-term recall in a developing country. *The Journal of Human Resources*, *36*, 593–625.

Beckett, M., Weinstein, M., Goldman, N., and Yu-Hsuan, L. (2000). Do health interview surveys yield reliable data on chronic illness among older respondents? *American Journal of Epidemiology*, *151*, 315–323.

Darwin, C. (1859). *The Origin of the Species*. New York: Modern Library.

Eaton, W. W., Holzer, C. E., von Korff, M., Anthony, J. C., Helzer, J. E., George, L., Burnam, M. A., Body, J., Kessler, L. G., and Lock, B. Z. (1984). The design of the Epidemiologic Catchment Area Surveys. *Archives of General Psychiatry*, *41*, 942–948.

Elder, G. H., Jr. and O'Rand, A. M. (1995). Adults living in a changing society. In K. S. Cook, G. A. Fine, and J. S. House (eds), *Sociological Perspectives on Social Psychology*, Boston: Allyn & Bacon, pp. 452–475.

Fowler, F. J. (1998). Design and evaluation of survey questions. In L. Bickman and D. J. Rog (eds), *Handbook of Applied Social Research Methods*, Thousand Oaks, CA: Sage, pp. 343–374.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., and Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. In C. C. Clogg (ed.), *Sociological Methodology 1988*, Washington, DC: American Sociological Association, pp. 37–68.

Henry, B., Moffitt, T. E., Caspi, A., Langley, J., and Silva, P. A. (1994). On the "remembrance of things past": A longitudinal evaluation of the retrospective method. *Psychological Assessment, 2*, 92–101.

Hirschi, T. (1969). *Causes of Delinquency*. Berkeley: University of California Press.

Menard, S. and Elliott, D. S. (1990). Longitudinal and cross-sectional data collection and analysis in the study of crime and delinquency. *Justice Quarterly*, *7*, 11–55.

Peters, H. E. (1988). Retrospective versus panel data in analyzing lifecycle events. *The Journal of Human Resources*, *23*, 488–513.

Pierret, C. R. (2001). Event history data and survey recall: An analysis of the National Longitudinal Survey of Youth 1979 recall experiment. *The Journal of Human Resources*, *36*, 439–466.

Radke-Yarrow, M., Campbell, J. D., and Burton, R.V. (1970). Recollections of childhood: A study of the retrospective method. In M. C. Templin (ed.), *Monographs for the Society for Research in Child Development*, *35* (5, Serial No. 138).

Rutter, M., Maughan, B., Pickles, A., and Simonoff, E. (1998). Retrospective recall recalled. In Robert B. Cairns, Lars R. Bergman, and Jerome Kagan (eds), *Methods and Models for Studying the Individual: Essays in Honor of Marian Radke-Yarrow*, Thousand Oaks, CA: Sage, pp. 219–243.

Scott, J. and Alwin, D. (1998). Retrospective versus prospective measurement of life histories in longitudinal research. In Janet Z. Giele and Glen H. Elder, Jr. (eds), *Methods of Life Course Research: Qualitative and Quantitative Approaches*, Thousand Oaks, CA: Ages Publications, pp. 98–127.

Taris, T. W. (2000). *A Primer in Longitudinal Data Analysis*. Thousand Oaks, CA: Sage.

Tulving, E. (1983). *Elements of Episodic Memory*. New York: Oxford University Press.

Weinfurt, K. P. and Bush, P.J. (1996). Contradictory subject response in longitudinal research. *Journal of Studies on Alcohol*, *57*, 273–282.

This page intentionally left blank

Chapter 8

A review and summary of studies on panel conditioning

David Cantor

1 Introduction

An important feature of a longitudinal survey is the possibility of panel conditioning. This refers to when participation in earlier waves of the panel affects the responses in subsequent waves. If there are significant conditioning effects, the utility of a longitudinal survey to measure change is compromised. The purpose of this chapter is to provide a review of the literature on conditioning, while trying to tie it together around three general goals. One goal is to describe the different types of conditioning that have been observed. As will be noted below, conditioning has been identified for almost all survey phenomena, including behavioral and attitudinal or opinion data. The review will provide examples of each of these. The second goal is to provide possible explanations for conditioning. This is important so that designers and analysts have a way to interpret and, possibly adjust, data when conducting analysis of panel data. The third goal is to provide the reader with the size of observed effects. It is one thing to say that conditioning effects exist, but like any other type of measurement error, it is important to understand how large these effects might be when making judgements about design and interpretation.

2 Theoretical and analytic principles

One reason why there is not a great deal known about panel conditioning is that it poses a number of analytic challenges. Unless one conducts a study specially designed to address conditioning, it is difficult to unequivocally state when it is occurring and why.

2.1 Theoretical distinctions

In their extensive review of different explanations of conditioning, Waterton and Lievesley (1989) enumerate six reasons why conditioning might occur:

1. Changing behavior or attitudes by raising consciousness.
2. Freezing attitudes.
3. More honest reporting of socially desirable behavior.
4. Improved understanding of the interviewing rules.

5. Higher motivation.
6. Lower motivation.

These distinctions are theoretically useful. They provide a framework for designing and analyzing panel data that exhibit conditioning effects. If a prior interview changes the phenomenon being measured (e.g., 1 or 2 above), then the result is partly a function of the measurement process, rather than something occurring in the population of interest. Adjustment of later waves of a panel or collecting alternative measures of change would need to be considered. If changes in the response process (points 3–6 above) is occurring, then it is critical to understand which explanation holds. If conditioning leads to higher motivation, less social desirability bias or greater understanding of the response task, then later waves of a panel are best. On the other hand, if conditioning is related to less desirable respondent behavior, such as decreased motivation or avoidance of additional burden, then responses to later waves of a panel have higher error.

Conditioning is a concern because it confounds measurement error with real change. Partly for this reason, rotating panel designs (Kalton and Citro, 1993) are typically structured to be balanced. This means that panels contributing to an estimate at a point in time are at the same degree of maturity. An assumption made when analyzing changes between periods for a rotating panel is that conditioning effects do not change over time and they are additive (e.g., rather than multiplicative). If this is true, then the change estimates across periods are unbiased.

2.2 Analysis requirements

One reason that the measurement of conditioning is not common is that it takes relatively elaborate designs to isolate the effects. Observation of changes in measures over time by itself is not evidence of conditioning because it confounds real change with conditioning. A common analytic strategy to isolate a conditioning effect is to compare two groups which have been exposed to the survey a different number of times, but have been interviewed during approximately the same calendar period. Differences between the groups should then theoretically be due to conditioning. One way to make this type of comparison is with a rotating panel design, where at each wave of a panel, there is a new sample being interviewed for the first time. Similar, but less ideal, designs compare panels to other cross-sectional surveys that are conducted during the same time period. For example, Wilson and Howell (2005) compare the trend in measures in the prevalence of arthritis for a panel to trends from a continuing cross-sectional survey (the National Health Interview Survey).

Nonresponse is a concern when assessing possible conditioning effects. Inevitably, data used from later panel waves will be subject to higher levels of nonresponse because of the cumulative effects of attrition. Studies that estimate conditioning effects typically implement a nonresponse adjustment. Some studies also try to simulate changes that assess whether those that drop out, given continuance of “no change” for the nonrespondents over waves, would affect conclusions about conditioning.

To rigorously assess the reasons behind conditioning, it would also be necessary to control for interviewer behavior and changes in interviewers between waves (Van Der Zouwen and Van Tilburg, 2001; O’Muircheartaigh, 1989). O’Muircheartaigh (1989) argues that to pinpoint the sources of conditioning one would need an experimental design that was specifically set up to tease out all of these effects (interviewers; respondents; interviewer–respondent interactions). Needless to say, this type of design has not been implemented. Nonetheless, it is still useful to review evidence of conditioning and assess how it fits within the processes Waterton and Lievesley (1989) enumerated. This should provide some perspective on the types of effects that have been found and how they might fit within the above theoretical framework.

3 Conditioning and changes in behavior

Several studies have found evidence of conditioning that leads to changes in behavior. The clearest example of this are studies of the electorate and voting behavior (Kraut and McConahay, 1973; Yalch, 1976; Traugott and Katosh, 1979), which have found that participation in a pre-election survey motivates respondents to subsequently vote. Clausen (1968) reports on a study of persons who participated in a pre-election survey in 1964. Voting information for those that participated in the pre-election survey were retrieved and compared to the proportion that voted in the general population. The analysis compares estimates from respondents on a panel survey who were interviewed prior to and after the 1964 election to estimates from one-time surveys that only interviewed after the election. A series of adjustments are made to the panel survey to make the population comparable to the other surveys, as well as trying to account for differences in nonresponse patterns. After adjustments, there remained a significant difference between the post-election estimates from the panel design and the cross-sectional survey. Up to 7 percentage points were attributed to the pre-election interview stimulating turnout for the actual election. Traugott and Katosh (1979) replicated these results and found a larger stimulus effect.

Wilson and Howell (2005) compare the prevalence of reports of arthritis in later panels of the Health and Retirement Survey (HRS) to cross-sectional estimates from the National Health Interview Survey (NHIS). The HRS rate of 371 per thousand in 1992 goes up to 415 per thousand in 1996, which is statistically significant. The NHIS rate for this time period stays virtually flat (296 to 304, not statistically significant). There are important differences between the two surveys. The questions used in the surveys are not identical and the surveys provide different contexts and ordering of survey topics. In addition, the HRS has a lower response rate at both the initial contact and from attrition over the panel. These differences are partly reflected by a prevalence rate for arthritis on the HRS that is 25% higher than on the NHIS. However, even after the authors attempt to adjust for differences in response rates, the divergent trends remain. In further support for the conditioning hypothesis, it is shown that a supplemental HRS sample introduced in 1998 exhibits similar prevalence rates as the first wave in 1992. As with the 1992 panel, the trend for the 1998 panel exhibits an upward trend in later panel interviews.

The authors discuss several possible explanations for the differences in trends between the NHIS and the HRS. The one they seem to think is best supported by the evidence is that the early interviews make respondents more aware of the possibility that they may have arthritis. This results in proactive follow-up with their doctors and even perhaps more actively asking about the possibility they may have arthritis. This argument is made through a process of elimination. By trying to control for differences in design and nonresponse, the most logical explanation remaining is the conditioning hypothesis. Unfortunately, the authors do not test this hypothesis for other diseases or other questions that may be common across the two surveys.

Veroff et al. (1992) follow up work by Wilson et al. (1984) suggesting that asking respondents about their feelings about their marriage will have an effect on the quality of the marriage. This study arose out of a concern from their Human Subjects Review Board that intensive probing of respondents about their marital relationship may raise concerns that may not have been brought up if the interview never took place. They test this by running two parallel panels. One panel consists of an intensive set of interviews with both the husband and wife about their relationship over a four-year period. Four relatively intense interviews were completed. The control group had fewer, and shorter, interviews over the four-year period. They found that the panel receiving more intensive interviewing had greater variance in their marital attitudes after the second interview. After the fourth year, they also found that for certain race–gender groups, the experimental group had more positive attitudes about their marriage, as measured by different scales on marital stability. For example, Blacks seemed to exhibit a much larger conditioning effect than Whites. While this study found some evidence of conditioning, the results were not consistent across groups. In addition, the study does not have an external measure of the quality of the marriage. They have only what respondents reported on the items within the survey.

Other evidence of behavior change is found by Battaglia et al. (1996). These authors hypothesized that asking mothers about the immunization status of their children will lead to respondents getting their babies vaccinated after the interview. The study was not a panel, but one which collected immunization records from respondents. This allowed the researchers to examine whether respondents vaccinated their children after the interview. They found that among those that reported their children as not having up-to-date vaccinations, 9.2% got at least one vaccination within 90 days after the interview. They conclude that for a panel survey, this would introduce an upward bias in prevalence estimates of approximately 2 percentage points (60% vs. 62%). As noted by the authors, this is an overestimate of the effect of conditioning, since it does not account for the natural growth of the percent of children that are vaccinated as they get older.

One theme that runs through many of the studies on conditioning is that respondents who are the least committed or certain about the outcome of interest will be the most subject to a conditioning effect. With respect to changing behaviors, the interview serves as a stimulus to take some action. For example, for voting it seems to serve to significantly raise respondent's propensity to vote. Consistent with this idea, Clausen finds that there seemed to be a bigger effect on turnout among Whites than for Blacks. This is attributed to the greater initial interest by Blacks in the election. For medical information, such as arthritis or immunizations, the interview may make respondents more aware of something that they feel they should be aware of. If this is true, then those most aware of the condition and/or the most able to act will be the least likely to change their behavior. Battaglia et al. (1996) found some evidence that those in lower income groups and with the least education were the most likely to get vaccinations after the interview. One might expect this assuming that these population groups are the least aware or least likely to get a vaccination.

One would expect that behaviors that are difficult or expensive to engage in will be least likely to be subject to this type of conditioning effect. For example, one would not expect making major consumer purchases to be influenced by an interview. Similarly, events that are not directly under the control of the respondent, such as victimization or being hospitalized, should not be influenced by an interview. Consequently, observed conditioning effects for these types of phenomena are likely to be the result of changes in the response process, rather than actual changes in behavior. In the next section we review evidence of conditioning for these types of surveys.

4 Conditioning and changes in the process for reporting behaviors

Some of the earliest evidence of conditioning came from consumer panels that asked about purchases (Prais, 1958; Ferber, 1953; Ehrenberg, 1960). Neter and Waksberg (1964a) report one of the first studies using a rotating panel. They compare the second and third interviews of the panel.[1] They find a significant drop of 9% in the reporting of household maintenance and repairs between the second and third interviews (173 million to 157 million). The drop was more pronounced for jobs of less than $20 value and for designated respondents such as the wife or any knowledgeable respondent. When the head of the household or a combination of the husband and the wife was selected, there was not a significant drop.

There are several other studies that have looked at consumer purchases. Silberstein and Jacobs (1989) report on the mean expenditures from the Consumer Expenditure Survey, an ongoing survey in the US that collects data for input into the Consumer Price Index. From Neter and Waksberg above, one might expect there would be a tendency for smaller purchases to be reported less often at later interviews. This should increase the mean expenditures at later times in sample. When looking at the mean expenditures for all types of purchases, Silberstein and Jacobs did not find any significant patterns. Significant changes were observed when looking at 7 of 17 more detailed expense classes. They find some evidence of decreased mean expenditures between the second and fifth interviews. However, this pattern was not consistent. For some expenditures, there was an increase rather than decrease. In addition, many of the statistically significant findings were not substantively very large.

Pennell and Lepkowski (1992) analyze reports of income from the Survey of Income and Program Participation (SIPP), a longitudinal survey. By comparing panels that overlap in time, they are able to compare data at a comparable calendar period from samples interviewed a different number of times. They examine income recipiency, income amounts, health insurance coverage, and labor force participation. They do not find evidence of a consistent effect of conditioning. The number of statistically significant differences in their analysis was less than what one would expect by chance (see also McCormick et al., 1992, for additional analyses of SIPP).

Frick et al. (2004) find the opposite in analysis of changes in the Gini coefficient measuring income inequality for the German Socio-Economic Panel (SOEP). Comparing the trends in a rotating panel design, it is found that the first interview of a newly introduced panel indicates higher income inequality than data from an older panel which is on later waves. Once respondents in the new panel are interviewed several times, however the measures converge with those provided by respondents to the longer running sample. After ruling out the possibility that this was due to either missing data or panel attrition, Frick et al. conclude that the difference was due to respondents getting better at the response task after the first several interviews.

The conclusion that accuracy of income reports increases over panel waves is supported by Rendtel et al. (2004) who analyze the European Community Household Panel (ECHP). Respondents from three countries participating in this panel (Denmark, Finland, and Sweden) have available income information from registers. By comparing the survey data to the register data, the analysis is able to assess whether changes at later panels also change in accuracy. Confirming the hypothesis by Frick et al. (2004), they do find accuracy to increase slightly over the first five waves of the survey. They also find indirect indicators of quality (e.g., missing data; use of estimation, rather than precise methods) to increase with panel waves.

Other studies examining reports of behaviors have looked at reports of medical conditions. This topic is similar to consumer behavior because it is relatively well defined and behaviorally-based. Corder and Horvitz (1989) compare quarterly estimates from a panel

[1] They do not use the first interview because it differs on key design features (e.g., length of recall period).

survey to estimates from a cross-sectional survey conducted during the same period. They examine a wide range of phenomena, including hospital discharges, first hospital visits, physicians visits, and reporting of acute conditions. They find no evidence of panel conditioning by comparing the trend in the panel data to the trend in the cross-sectional data.

Cohen and Burt (1985) find that collecting data more frequently on a panel led to significantly fewer medical events and medical expenditures being reported. This study compared these measures for a group of people interviewed five times to a group interviewed 4 times during the same 1-year period. The magnitude of the differences ranged from 7% to 16%. The authors argue that the data collection frequency led to reductions in reporting. This argument was supported by evidence that respondents reported more accurately in the four round group, as indicated by a greater concordance with provider records collected during a follow-up of providers for whom the respondent reported seeing.

While both expenditures and visits to the doctor are well-defined events, there are other behavioral phenomena that have more ambiguity associated with them. One might argue that conditioning may be greater for these types of events because respondents are more likely to use the first interview to learn about the objectives of the survey than may be the case for more straightforward types of behaviors. Respondents may become better prepared to answer questions, once they have been exposed to the entire questionnaire (Waterton and Lievesley, 1989; Cannell et al., 1981; Biderman and Cantor, 1984). An alternative argument is that respondents may figure out ways to reduce burden by saying they do not have the condition or were not engaged in the behavior, e.g., when an affirmative response to questions about the behavior leads to follow-up questions asking for details about the behavior.

One example of this is for reports of victimization. These events are subject to idiosyncratic definitions that are influenced by the cues provided to respondents (Cantor and Lynch, 2000). The National Crime Survey (NCS) has a rotating panel design and exhibits significant decreases in reports of victimization at later panel waves. The largest drop is between the first and second interview (30%–40%). This is followed by drops of around 15% in later waves of the panel (Cantor, 1989). The large drop between the first and second interview is explained by the fact that the reference period for the second interview is temporally bounded by the first interview. The first interview does not have this temporal bound. Bounding is thought to minimize telescoping events into the reference period (Neter and Waksberg, 1964b).

A related phenomenon is self-reported perpetration of criminal or delinquent behavior. Thornberry (1989) compared time trends for the National Youth Survey (NYS) to the Monitoring the Future (MTF). The NYS is a panel survey and MTF is a repeated cross-sectional survey. Significant differences in the trend of these data were found. The author concludes that this is the result of conditioning on the NYS. Menard and Elliott (1993) re-analyze the same data and argue that the comparisons made by Thornberry do not appropriately consider the differences in methodology (e.g., question wording, sample design). For example, the cross-sectional portion of the MTF is of high school seniors, whereas the NYS is a sample of youth between 11–24 (depending on the cohort). In this reanalysis, no significant evidence of conditioning for the NYS was found.

A third, relatively ambiguous set of behaviors is unemployment. The definition of unemployment depends on idiosyncratic definitions of labor force participation. Behavior related to “looking for work” might be difficult to define uniformly across respondents. Significant drops in measures of unemployment have been found in several different labor force surveys (Bailar, 1989; Ghangurde, 1982). Bailar compares different labor force statistics across the time in samples for the Current Population Survey (CPS). Rates of unemployment at the first interview are found to be 7% higher than the average over the life of the panel. There are also small, statistically insignificant, drops in the rate over the next two interviews.

For almost all of the above studies, significant effects take the form of a reduction in the reporting of the particular phenomena (e.g., expenditures, victimization, unemployment). A "burden" explanation for the drop is that respondents are avoiding the extra questions associated with reporting the behavior (e.g., unemployment). A second possibility is that respondents are better prepared or more motivated to answer questions after the initial interview. There isn't any direct evidence to address either explanation. Re-interview data from the CPS (Bailar, 1975) suggests that estimates of unemployment generally are too low. One might therefore assume that higher rates of unemployment represent more accurate information. It is the case, however, that there is substantial error associated with the re-interview data as well. Even if one accepts the re-interview as a less biased estimate, we are not aware of any data that breaks it out by time in sample. Furthermore, it is not entirely clear how respondents are avoiding burden when reporting activities related to being in or out of the labor force.

With respect to unemployment, a second possibility might be that the first interview is not bounded by any other interviews. Determination of unemployment status relies on a reference period of four weeks. It may be that some of the pattern is due to respondents telescoping from outside the four-week period. For example, it may be that the subject did look for a job three months ago, but the respondent is misdating it at the first interview. This may not occur for all of the other times in sample which have some type of bound by a previous interview. In addition, respondents may not quite understand the precision required by the time reference at the first interview. They may be relatively eager to respond in a positive way and may tend to include behavior from outside the reference period. Once completing the first interview, they become more knowledgeable about these information demands and are less likely to report labor force activity outside the reference period.

5 Conditioning and reports of attitudes, opinions and subjective phenomena

Attitudes, opinions, and subjective phenomena are less well defined than the behavioral phenomena above. Van der Zouwen and van Tilburg (2001) discuss panel conditioning of these types of reports in the context of trying to measure a latent characteristic. They make the distinction between interviews changing respondent attitudes or changing the way the questionnaire items measure the attitude. The former is equivalent to changing the latent trait that is being measured, similar to changing behavior as discussed in Section 3. The latter relates to changes in the measurement of the latent trait. While the distinction is important, no evidence has been generated that allows separating the two. Consequently, for purposes of discussion, we combine both in this section.

For behavioral questions such as victimization, reporting an event often leads to getting asked additional questions. For attitudes and opinions, no such branching is typically used. Perhaps for this reason, a number of researchers that find conditioning effects argue that data are improved because of a learning or motivation effect over panel waves. The empirical evidence of a panel effect for attitudes, opinions, and subjective reports is mixed. In some cases, there seem to be effects of conditioning on the expression of certain attitudes, while for other topics there are no effects found. Bridges et al. (1977) hypothesize that panel conditioning will occur for those topics that respondents perceive as being important, but are not particularly informed about. If respondents are not informed about a topic covered at an initial interview and they view it as important, they will form opinions that are expressed at the follow-up interviews. For attitudes that are already formed or refer to unimportant topics, no changes will occur.

To partially test this, they conducted an experiment looking at two different attitudes. At the first interview, one half of the sample was interviewed about their concern about cancer. The other half was interviewed on their concern about burglary prevention. Both groups were asked two general questions about their concern about good health and about crime. At the second interview, both groups were given identical questionnaires with items on both cancer and burglary. The results show that those interviewed about cancer in the first interview exhibited an increased concern about good health at the second interview. A similar pattern did not result for those initially interviewed about burglary. Those interviewed about burglary at the first interview did not show an increase in concern about crime in the second interview. The change in attitudes about general health ranged from 8% to 11%. They also find that a mailing about burglary prevention that was in between the two interviews increased the change in responses between interviews. Those that received the mailing were more likely to report concern for both health and crime at the second interview. This effect was approximately of the same magnitude of prior interviews (around 10%).

Bridges et al. account for this pattern by noting that respondents had more specific concerns about burglary than about cancer. Less concern is indicative of how well formed the attitude is in the respondent's mind. The less formed, the more likely the respondent will be to change their answer. The authors offer evidence of this by noting that at the initial interview, respondents showed higher concern about burglary than cancer. With respect to the mailing, the authors interpret this as evidence that providing more information about safety issues raised concerns on not only crime, but also general health issues.

A related explanation has been given to observed panel effects for life satisfaction. Analysis of the German Socio-Economic Panel (SOEP) finds that these measures decrease over time (Landua, 1991; Frick et al., 2004; Jürges, 2005). From data contained in Frick et al. (2004), approximately 20% more respondents select the top three scale points to the satisfaction question when interviewed the first time relative to those who had been asked in prior interviews (Table F-1 for 2002, balanced cross-section). Graphs in Jürges (2005) show a drop from 40% of respondents choosing the most satisfied category in 1984 to less than 5% 14 years later. This compares to a much smaller decrease for a repeated cross-sectional survey conducted during this same time period (about 20% to 13%). Frick et al. (2004) argue that the drop over the panel is due to respondents becoming more familiar with their satisfaction by thinking about the question prior to the second, third, etc . . . interviews (see also, Jürges, 2005).

This interpretation of a "learning effect" of repeated interviewing is viewed as reducing overall measurement error. However, no empirical data is provided which identifies this explanation over others that might attribute the drop as increasing error (e.g., due to changes in interviewers; increases in social desirability).

Social desirability has also been hypothesized as a cause of panel conditioning. Pevalin (2000) cites several studies that observed a decrease in mental health scores as measured by the General Health Questionnaires (GHQ). These studies conducted interviews over both a relatively short (e.g., multiple interviews within a one-year period) and long (once a year for multiple years) periods of time with both specialized and general population samples. None of these studies had a parallel group that controlled for trends over time or possible effects of aging. They also used different versions of the GHQ (12 item vs. 30 item vs. 60 item). A few of these studies did conduct more detailed diagnostic interviews as part of the protocol. These few studies found evidence that the GHQ lost sensitivity over time. The drop in the GHQ over time was attributed to an increasing desire to appear to the interviewer that they are mentally healthy.

Pevalin's own analysis of the British Household Panel Study (BHPS) compared the GHQ 12 item version to an annual cross-sectional survey done in Britain (Health Surveys for England – HSE). No evidence of conditioning was found. Unlike prior studies cited in the article, scores increased, rather than decreased, over time. The author hypothesizes that the difference from other studies could be the relatively long time interval between survey administrations for the HSE and BHPS. The large re-test effects found in other studies tended to occur when the administration of the survey was considerably shorter, with it being conducted multiple times in a year.

Sobol (1959) examines attitudes related to perceptions of individual economic well-being (e.g., better off than last year? Expect to be better next year?). The analysis initially finds a significant drop in these attitudes relative to a cross-sectional survey. However, this analysis finds that once controlling for attrition in the panel, the differences disappear. Those in lower income groups, renters and those not interested in the survey tended to drop out after the first interview. Once examining only those that stayed in the panel for the entire period, no trend in the measures was found.

Waterton and Lievesley (1989) test several different hypotheses on the causes of conditioning with the British Social Attitudes Panel. This was a small panel (approximately 800 at the first wave) that originated with the Social Attitudes Survey. By comparing measures from the panel to cross-sectional measures for a series of three different years, they test whether repeated interviewing changes attitudes by examining changes in political attitudes over the three waves of the study. They find changes do occur in the predicted direction. There are increases in political partisanship of approximately 7 percentage points over a 3-year period (51% to 58%). Similarly, they test whether conditioning occurs to reduce the social desirability bias over time by examining six different items on racial prejudice, social action (protesting an unjust law; break law to follow conscience), refusing an income question, propensity to break the law, and newspaper readership. They find ambiguous evidence for this, with three of the items being in the right direction, but only one of these being statistically significant. It should be noted, however, that the panel had relatively small sample sizes to detect rather low prevalence rates (e.g., 4.9% of the sample who admit to being racially prejudiced).

They also test for social desirability a second way by examining the inter-item correlations among related items. They hypothesize that the inter-item correlations should go down over time because respondents are not attempting to stay consistent within a questionnaire. They do not find any evidence supporting this hypothesis for any of the items tested.

Seemingly, large panel effects were found in a study related to a panel survey on young people's (age 16–24) intention to enlist in the military (Nieva et al., 1996). The Youth Attitude Tracking Survey (YATS) was an ongoing general population telephone survey that had a rotating panel design. The panel was set up to save on the costly process of screening for youth who were eligible for the interview. Comparison of the first interviews to those interviewed a second and third time found drops in the respondent's desire to enlist by as much as 33% (e.g., 20.7% to 14.7%). This study had significant panel attrition, as no allowance was made to track those who moved between panel waves. Approximately 40%–50% of the initial sample was not interviewed a second time. For this reason, the above results are confounded by nonresponse bias. Those who were most likely to drop out tended to be the most positive on the enlistment propensity measure. Nonetheless, using various nonresponse adjustment strategies, the significant drop in propensity was not dramatically affected.

The conflicting results across these studies points to the need to develop more detailed explanations for the occurrence of conditioning effects. Up to this point in time, there has been little in the way of validating the competing explanations discussed by Waterton and Lievesley (1989). Consequently, it is difficult to predict when conditioning effects might occur, either with respect to content area or survey design. In the summary below, we describe the different frameworks that might be used in trying to disentangle conditioning effects.

6 Summary and discussion

The results of selected studies of panel conditioning are summarized in Table 8.1. Studies on panel conditioning have found effects that should concern analysts of longitudinal data. Generally, the size of these effects range from 5% to 15%, with a few finding greater effects. There are a number of studies that have not found significant panel effects. There are also instances where the same phenomena are studied with different results (e.g., consumer purchases). The mixed results may be partly due to the complicated analytic requirements to identify the effects and to be able to pinpoint the causes. In order to identify conditioning effects, it is necessary to separate real change due to trends in the phenomena of interest, effects of panel attrition, and changes that occur across panels (e.g., interviewers and interviewer–respondent interactions).

We have organized the above discussion around a continuum with one end being behaviors that may be changed as a result of an interview, the middle being reports of behaviors that might be changed with no real change in behavior, and the other end being changing latent traits, such as attitudes, opinions, and subjective assessments. This continuum reflects the different content areas affected, as well as the way one might measure conditioning effects.

Several behaviors have been found to be affected by conditioning. In the context of voting, a pre-election interview increases the respondent's sense of civic duty, which directly affects their voting propensity. It may also increase the perceived obligation to vote. With respect to medical conditions, the interview puts to the forefront issues that are of important personal concern to the respondent. For example, the review above cited the case of asking about arthritis where it was hypothesized that conditioning may have motivated respondents to see a doctor and get it diagnosed (e.g., when someone has been experiencing pain, but had no explanation). A similar motivation was suggested when asking about vaccinations.

The evidence and analyses are not definitive. It is difficult to predict when conditioning will change behaviors. Asking about medical conditions, for example, does not have the same effect for all respondents and all medical issues. For example, as noted above, Cohen and Burt (1985) report decreases, rather than increases, in reports of health care utilization and expenditures with greater interviewing frequency. Similarly, Corder and Horvitz (1989) do not report any indications of conditioning for reporting medical expenditures. There is no apparent change for consumer expenditures. What may distinguish the studies that involve changing behavior is that they involve behaviors that are relatively easy to engage in (e.g., voting). In addition, changes may only occur for populations that are particularly close or affected by the topic. For example, the study by

Table 8.1 Summary of selected studies testing for panel conditioning by type of measure

	Variable	*Analysis*	*Effect size**
Change Behavior			
Clausen, 1968	Voting	Compared to cross-section Voting records	8%
Traugott and Katosh, 1979	Voting	Compared to cross-section	9%
Wilson and Howell, 2005	Arthritis	Compared to cross-section	12%
Veroff, et al., 1992	Marital stability	Compared to group with less intense interviews	14% to 100% Blacks only
Battaglia, 1996	Immunization	Checked records after interview	3%
Reporting Behaviors			
Neter and Waksberg, 1964a	Expenditures on household maintainence	Rotating panel design	9%
Silberstein and Jacobs, 1989	Consumer expenditures	Rotating panel design	ns
Pennell and Lepkowski, 1992	Income	Rotating panel design	ns
Frick, et al., 2004	Gini coefficient for reports of income	Rotating panel design	can't compute
Corder and Horvitz, 1989	Hospital discharges Hospital visits Physicians' visits Acute conditions	Compared to cross-section	ns
Cohen and Burt, 1985	Medical events Medical expenditures	Compared 5 rounds of interviews to 4 rounds of interviews	7%–16%
Cantor, 1989	Victimization	Rotating panel design	15%
Menard and Elliott 1993	Criminal offending	Compared to cross-section	ns
Bailar, 1989	Unemployment	Rotating panel design	7%
Reports of attitudes, opinions and subjective phenomena			
Bridges, et al., 1977	Concern with health Concern about crime	Experimented with inclusion of question in one group	8%–11% for health ns for crime
Frick, et al., 2004	Life satisfaction	Compared to cross-section	20%
Pevalin, 2000	Mental health	Compared to cross-section	ns
Sobol, 1959	Perceptions of economic well-being	Compared to cross-section	ns
Waterton and Lievesley, 1989	Political attitudes Racial predjudice Social action Propensity to break law	Compared to cross-section	13% for political attitudes ns for other attitudes
Nieva, et al., 1996	Propensity to join military	Rotating panel design	33%

*Computed as [(C-P)/C] × 100 where C = cross-sectional estimate, P = panel estimate; ns = not statistically significant.

Wilson and Howell (2005) studied adults aged 55–56 years old. This is an age when people may be just getting concerned with arthritis. For vaccinations, new mothers may be the most susceptible to a conditioning effect. The fact that Battaglia et al. (1996) found this effect was slightly greater for the most economically marginalized is consistent with this idea.

Explaining why reports of behaviors are affected by conditioning is more complicated. When effects are observed, the pattern is for quantities or rates of behaviors to decrease over the life of the panel. This pattern is not observed for all those discussed above. In the case of reporting income, consumer expenditures, and medical events, there were studies that found both negative and positive evidence of conditioning. This ambiguity may be related to factors that are not easily controlled. For example, the spacing between interviews may have significant effects on conditioning. Shorter time periods between interviews should increase effects because it may increase the tendency for respondents to remember what happened during the previous interview (e.g., Kalton et al., 1989). Interviewer effects are uncontrolled in all of the studies reviewed above, except for one.

Two explanations have been offered to explain why conditioning may affect reports of behaviors. One is that respondents learn that affirmative answers lead to additional questions. They then avoid answering affirmatively in the future. A common assumption is that respondents may actively avoid questions if they understand the implications of answering in a particular way. A common model that summarizes this hypothesis is that respondents tend to satisfice (Krosnick, 1991) more in later panels of a survey. Once the uniqueness of the survey task wears off, respondents view the task more routinely and may not exert as much effort at hard tasks, such as recall over extended periods of time.

There is very little data to assess the burden explanation. We could not find studies that show that respondents avoid certain responses because they believe it will increase the burden on the survey. One example that we could find is a study by Turner (1984), who reports that fewer crime incidents are reported when respondents are asked detailed questions right after they answer affirmatively to each screening question. Even in this case, however, it isn't clear if respondents are reacting to the burden or are using the detailed questioning to get a better idea of the type of phenomena of interest to the study. If burden does play a role in responding, it isn't clear the conditions under which it occurs. It isn't clear how it applies across panel waves, for example. Similarly, how do perceptions of burden relate to short versus long interviews? One might expect, for example, the effect to be a function of the overall burden and sensitivity of the survey. Long surveys may tempt respondents more to avoid burden than relatively short surveys.

A view that burden is a key driver differs sharply from the view that respondents are generally motivated to respond, learn about the survey at the first administration and use this knowledge when answering questions at subsequent waves. Under a learning hypothesis, the respondent is seen as initially being very anxious to provide responses that might be of use to the study. It is partly through this motivation that they may actually over-report when initially coming up with a response. After they complete the survey the first or second time, their understanding and response formulation are adjusted to reflect this additional thinking (e.g., Biderman et al., 1986; Cannell et al., 1981).

One way to distinguish between these two hypotheses would be to examine measures of data quality over the life of a panel survey. The burden hypothesis would predict that data quality decreases, while the learning/motivation hypothesis predicts that it improves. While theoretically this question could be addressed with standard survey methods, there are very few studies that have actually done so. Bailar (1989) reports on several studies that seem to show a decrease in the quality of reporting expenditures (Pearl, 1979) and illnesses (Mooney, 1962). As noted above, Cohen and Burt (1985) also show a decrease in the correlation between reports and records for provider visits. Contrary evidence is reported for the reporting of income by Rendtel et al. (2004) who find improved data quality for reports of income when checking against records.

One might also expect variation in panel effects across different types of behavioral phenomena. Under a learning/motivation model, behaviors that are relatively well defined may not greatly benefit from a prior interview. Thus phenomena like hospitalizations and income may not exhibit significant conditioning effects. On the other hand, for phenomena that are more difficult to define, there would be larger conditioning effects because the survey provides a great deal of information that the respondent can use the next time. The evidence on this is decidedly mixed across studies, as evidenced by conflicting effects for the same phenomena (e.g., consumer expenditures).

This suggests that the reasons for conditioning are more complicated than the above explanations would imply. Recent cognitive and social psychological theories of the response process do view respondents more along a continuum with respect to their approaches to the response task (Tourangeau et al., 2000, pp. 1–22). Some respondents will be anxious to exert effort to do whatever it takes to complete the task. This would include trying very hard to understand the meaning of the questions, recalling information from memory (or taking actual or mental notes between panel waves) and fitting their responses within the structure of the survey. Others will exert as little effort as possible, while many will be somewhere in the middle. It is likely that survey characteristics (e.g., difficulty of the task), the interest respondents have in the topics, and the respondent's cognitive abilities, among other things, would influence the extent that the burden or learning/motivation hypotheses might be true.

Similarly, the influence interviewers might have on how well respondents carry out the task may be very important. Interviewers may change between waves and collect the data in different ways (e.g., how they probe), which may lead to interviewer effects. Conversely, the same interviewer may probe in ways that lead to a systematic change over time. For example, van der Zouwen and van Tilberg (2001) found from recording of interviews a tendency for some interviewers to use data from the prior interview to shape the answers at subsequent waves. They found this led to a decrease in the number of reports of interest (network size). In addition, they argue that the use of these data by the interviewer are shaped by a desire to keep the interview as short as possible. Rather than an effect that is due to respondent's avoiding burden, these data indicate that it is the interviewer who is shaping the responses, through use of data from the previous interview, to reduce the number of reports.

The length of a survey is not clearly linked to how one answers attitudinal questions. Therefore when moving to explaining conditioning for latent characteristics, such as attitudes, burden is not clearly relevant. Perhaps for this reason, the causal effects related to conditioning of attitudinal data is not as much a subject for debate in the literature. Most of those who find conditioning effects on attitudes or opinions attribute changes to a learning effect (Sturgis and Allum, 2004; Waterton and Lievesley, 1989). For example, Sturgis and Allum (2004) argue that exposure to attitude questions at the initial interview stimulates respondents to think more about the topic. This crystallizes opinions that had not been considered up to that point. The stimulus strengthens or even changes the attitude measured at the second and subsequent interviews (see also Frick et al., 2004). Interestingly, these discussions do not make much of the difference between a change in the attitude or in how the attitude is reported.

As in the case of conditioning and reports of behaviors, there is very little evidence that would support a learning/motivation hypothesis. However, theories of the response process as they apply to the context and order effects for attitude questions (Sudman et al., 1996; Tourangeau et al., 2000) are relevant when predicting conditioning effects. For example, the inclusion/exclusion model views context effects as a function of what information is accessible in memory to construct an answer. Preexisting knowledge used to answer a question is called "chronically accessible." Context effects occur when information temporarily added from exposure to a prior question is more influential than chronically accessible information. Context effects are difficult to predict because they depend on not only the questions, but the relationship between chronically and temporarily accessible information.

Of course conditioning is different from context effects. The latter are explained by referring to temporarily accessible memory from immediately preceding questions. They are less likely to occur as the spacing between similar questions increases. Nonetheless, the basic idea of prior questions serving as stimuli for generating information relevant to formulating responses is how the learning model of conditioning is proposed to operate. In this case, the initial exposure to a question adds to chronically accessible information that is used when answering items at subsequent interviews. Information may also be added through any stimulus effect the initial interview has on respondents thinking about the attitude object. When the re-interview occurs, respondents draw on these data and may form a different judgement when compared to the prior interview. Different information is available for use when making the judgement. Note that this explanation does not necessarily rely on a respondent explicitly remembering that they answered the question the last time. It only relies on the assumption that the process of answering the question the first time changes what is eventually accessible in memory the next time the question is asked.

The review in this chapter points to evidence of conditioning effects for questions about behavior and latent characteristics. To fully understand the boundaries around when these effects occur, it will be necessary to design studies that can make important distinctions between the competing explanations. Drawing on more sophisticated models of the survey process, as well as simulating the conditions surrounding panel designs, may be the most promising way to make progress in understanding how these effects influence the accuracy of panel estimates.

References

Bailar, B. (1989). Information needs, surveys, and measurement errors. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys*, pp. 1–24. New York: Wiley.

Bailar, B. A (1975). The effects of rotation group bias in estimates from panel surveys. *Journal of the American Statistical Association,* 70: 23–30.

Battaglia, M., Zell, E. R. and Ching, P.L.Y.H. (1996). Can participating in a panel sample introduce bias into trend estimates? *1996 Proceedings of Survey Research Methods Section of the American Statistical Association*, pp. 1010–1013.

Biderman, A. D. and Cantor, D. (1984). A longitudinal analysis of bounding, respondent conditioning and mobility as sources of panel bias in the National Crime Survey. *Proceedings of the Survey Research Section of the American Statistical Association*, pp. 708–713.

Biderman, A. D., Cantor, D., Lynch, J. P. and Martin, E. (1986). *Final Report of Research and Development for the Redesign of the National Crime Survey*. Washington D.C: Bureau of Social Science Research.

Bridges, R. G., Reeder, L. G., Kanouse, D., Kinder, D. R., Nagy, V. T. and Judd, C. M. (1977). Interviewing changes attitudes—sometimes. *Public Opinion Quarterly,* 41: 56–74.

Cannell, C., Miller, P. V. and Oksenberg, L. (1981). Research in interviewing techniques. In

S. Leinhardt (ed.), *Sociological Methodology 1981* pp. 389–437. San Francisco: Jossey-Bass.

Cantor, D. (1989). Substantive implications of selected operational longitudinal design features: The National Crime Survey as a case study. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys*, pp. 25–51. New York: Wiley.

Cantor, D. and Lynch, J. P. (2000). Self-report surveys as measures of crime and criminal victimization. In D. Duffee, D. McDowall, B. Ostrom, R. D. Crutchfield, S. D. Mastrofski and L. G. Mazerolle (eds), *Measurement and Analysis of Crime and Justice*, pp. 85–138. NCJ 182411, Washington D.C: US Department of Justice.

Clausen, A. (1968). Response validity: vote report. *Public Opinion Quarterly*, 41: 56–64.

Cohen, S. B. and Burt, V. L. (1985). Data collection frequency effect in the National Medical Care Utilization and Expenditure Survey. *Journal of Economic and Social Measurement*, 13: 125–151.

Corder, L. S. and Horvitz, D. G. (1989). Panel effects in the National Medical Care Utilization and Expenditure Survey. In D. Kasprzyk, G. J. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys*, pp. 304–318. New York: Wiley.

Ehrenberg, A. S. C. (1960). A study of some potential biases in the operation of a consumer panel. *Applied Statistics*, 9: 20–27.

Ferber, R. (1953). Observations on a consumer panel operation. *Journal of Marketing*, 17: 246–259.

Frick, Joachim R., Goebel, Jan, Schechtman, Edna, Wagner, Gert G. and Yitzhaki, Shlomo (2004). Using analysis of Gini (ANoGi) for detecting whether two subsamples represent the same universe: the German Socio-Economic Panel Study (SOEP) experience. Institute for the Study of Labor (IZA) Discussion Paper No. 1049, Bonn, Germany. Available at www.iza.org.

Ghangurde, P. D. (1982). Rotation group bias in the LFS estimates. *Survey Methodology*, 8: 86–101.

Jürges, H. (2005). First reply effects (aka repeated measurement effects). Presentation at the DIW Workshop, Methodology and Measurement of Subjective Variables, Berlin, May 2005.

Kalton, G. and Citro, C. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology*, 19(2): 205–215.

Kalton, G., Kasprzyk, D. and McMillen, D. B. (1989). Nonsampling errors in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys*, pp. 249–270. New York: Wiley.

Kraut, R. E. and McConahay, J. G. (1973). How being interviewed affects voting: an experiment. *Public Opinion Quarterly*, 16: 381–398.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5: 213–236.

Landua, D. (1991). An attempt to classify satisfaction changes: Methodological and content aspects of a longitudinal problem. *Social Indicators Research*, 26: 221–241.

McCormick, M. K., Butler, D. M. and Singh, R. P. (1992). Investigating time in sample effect for the Survey of Income and Program Participation. Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 554–559.

Menard, S. and Elliott, D. S. (1993). Data set comparability and short-term trends in crime and delinquency. *Journal of Criminal Justice*, 21: 433–445.

Mooney, H. W. (1986). Methodology in two California health surveys. *Public Health Monograph, No. 70*.

Neter, J. and Waksberg, J. (1964a). Conditioning effects from repeated household interviews. *Journal of Marketing*, 28: 51–56.

Neter, J. and Waksberg, J. (1964b). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Association*, 59: 18–55.

Nieva, V., Morgenstein, D., Wilson, M., Cantor, D., Chu, A., Hintze, W., Lehnus, J. and Nathan, G. (1996). Youth Attitude Tracking Study: Panel effects in enlistment propensity. Report submitted to the Defense Manpower Data Center, US Department of Defense, under contract number MDA903-90-0236.

O'Muircheartaigh, C. (1989). Sources of nonsampling error: discussion. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys*, pp. 271–288. New York: Wiley.

Pearl, R. B. (1979). Re-evaluation of the 1972–73 US Consumer Expenditure Survey. Technical Paper No. 11. Washington D.C.: US Bureau of the Census.

Pennell, S. and Lepkowski, J. M. (1992). Panel conditioning effects in the survey of income and program participation. Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 566–571.

Pevalin, D. J. (2000). Multiple applications of the GHQ-12 in a general population sample: An investigation of long-term retest effects. *Social Psychiatry and Psychiatric Epidemiology*, 35: 508–512.

Prais, S. J. (1958). Some problems in the measurement of price changes with special reference to cost of living. *Journal of the Royal Statistical Society*, Series A., Part 3: 312–332.

Rendtel, U., Nordberg, L., Jäntti, M., Hanisch, J. and Basic, E. (2004). Report on quality of income data. Chintex Working Paper #21, Work-package 5. Retrieved from website http://www.destatis.de/chintex/proj_des/wp_5.htm

Silberstein, A. R. and Jacobs, C. A. (1989). Symptoms of repeated interview effects in the Consumer Expenditure Interview Survey. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys*, pp. 289–303. New York: Wiley.

Sobol, M. G. (1959). Panel mortality and panel bias. *Journal of the American Statistical Association*, 54: 52–68.

Sturgis, P. and Allum, N. (2004). Panel conditioning and scale reliability, evidence from the British Household Panel Study. Paper presented at the Annual Meeting of the American Association for Public Opinion Research.

Sudman, S., Bradburn, N. and Schwarz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Thornberry, T. P. (1989). Panel effects and the use of self-reported measures of delinquency in longitudinal studies. In M. W. Klein (ed.), *Cross-National Research in Self-Reported Crime and Delinquency*, pp. 347–369. Dordrecht, Netherlands: Kluwer Academic Publishers.

Tourangeau, R., Rips, L. J. and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Traugott, M. W. and Katosh, J. P. (1979). Response validity in surveys of voting behavior. *Public Opinion Quarterly*, 43: 359–377.

Turner, A. G. (1984). The effect of memory bias on the design of the National Crime Survey. In *The National Crime Survey: Workshop Papers, Vol. 2, Methodological Studies*. Washington DC: US Department of Justice, Bureau of Justice Statistics.

van der Zouwen, J. and van Tilburg, T. (2001). Reactivity in panel studies and its consequences for testing causal hypotheses. *Sociological Methods and Research*, 30: 35–56.

Veroff, J., Hatchett, S. and Douvan, E. (1992). Consequences of participating in a longitudinal study of marriage. *Public Opinion Quarterly*, 56: 315–327.

Waterton, J. and Lievesley, D. (1989). Evidence of conditioning effects in the British Social Attitudes Panel Survey. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys*, pp. 319–339. New York: Wiley.

Wilson, S. E. and Howell, B. J. (2005). Do panel surveys make people sick? US arthritis trends in the Health and Retirement Survey. *Social Science and Medicine*, 60: 2623–2627.

Wilson, T. D., Dunn, D. S., Bybee, J. A., Hyman, D. B. and Rotondo, J. A. (1984). Effects of analyzing reasons on attitude-behavior consistency. *Journal of Personality and Social Psychology*, 47: 5–16.

Yalch, R. F. (1976). Pre-election interview effects on voter turnout. *Public Opinion Quarterly*, 40: 331–336.

Chapter 9

Reliability issues in longitudinal research

Toon W. Taris

1 Introduction

The reliability of a concept refers to the degree to which consecutive measurements of this concept yield the same result, given that the underlying score on the concept has not changed. In this chapter we deal with various issues relating to the reliability of measures across time. In Section 3 we discuss a procedure to establish whether the associations among the items of repeatedly measured multi-item measures has remained the same across time. If so, the reliability as well as the meaning of this concept has not changed across time—which is imperative if one is to make across-time comparisons. Then we discuss the implications of measurement unreliability for across-time comparisons, focusing on issues such as regression to the mean (the tendency for subjects with extreme scores to obtain scores that are closer to the group average at subsequent measurements) (Section 4), the reliability of change scores, and the regression fallacy, i.e., when change in the criterion variables that is due to measurement unreliability is attributed to the independent study variables (both in Section 5).

2 Reliability issues in longitudinal research

Classical test theory defines reliability in terms of the degree to which consecutive measurements of a particular concept yield the same result, given that the underlying score on the concept has not changed. The scores on any given measure are presumed to reflect (1) the score on the concept of interest (the *true score*), and (2) *error*, which may include *bias* (systematic error, e.g., a scale that consistently underestimates weight by 5 kilograms, and which reduces validity but not reliability) and *random error* in which the error is not consistently one of overestimation or underestimation (and which reduces both reliability and validity). For example, a person's score on an intelligence test may reflect his cognitive ability, but perhaps also administration mode-related factors (e.g., experience with computer-administered tests), contextual factors (noise, temperature), or familiarity with intelligence tests. Researchers usually aim to maximize true score variance, relative to error variance;

measures that largely reflect error rather than respondents' true scores are useless as they cannot tell us much about the phenomena that interest us nor about their interrelations. Thus, researchers must have some idea of the reliability of their measures, as this directly affects the strength of the inferences that can be made on the basis of their study.

Some phenomena in social, educational, and behavioral research can easily be measured reliably. For example, concepts such as participant gender, year of birth, or level of education can be measured accurately with a single item. For such concepts the true score variance/error variance ratio will be acceptable, even if just a single item is used. However, in other cases this ratio will be less favorable, and in that case multi-item measurement is needed to obtain reliable estimates of the participants' true scores on the phenomena of interest. For instance, personality traits, attitudes, intentions, behaviors, and well-being are often measured using multi-item measures. The scores on the separate items of these measures may contain a large error component, but as these errors are presumed to be due to random factors, these should largely cancel each other out. Multi-item measures will therefore give a more reliable indication of the participants' true scores than single-item measures.

Reliability estimation

In longitudinal research, two forms of reliability estimation are especially important: (1) coefficients of internal consistency (taking into account the degree to which the components of the test are correlated) and (2) coefficients of stability (test-retest reliability). A third form of reliability estimation focuses on equivalence, i.e., the degree to which measures that are presumed to measure the same construct are correlated.

1. *Internal consistency estimation* starts from the assumption that the items belonging to a particular scale should be considered replications, so that similar responses should be given to any pair of them. Less than perfect correlation indicates that the items do not tap precisely the same facet or level of the underlying construct. Cronbach's alpha coefficient is currently the most widely used example of this type of reliability estimate and focuses on the ratio of true score variance and error variance. According to Nunnally (1978), alpha should be .60 at minimum (and preferably exceed .70) to be acceptable, meaning that at the very least 37.5% of the variance on any given measure should reflect true score variance. Estimation of the internal consistency of measures does not necessarily require a longitudinal design; alpha can be estimated if there are at least two measures (items) for the concepts of interest, which can easily be achieved within most cross-sectional studies.
2. Estimation of *test-retest stability* obviously requires some time interval to pass between the test and the retest, and therefore always calls for a longitudinal design. As Robert Guion (2002) notes, the correct length of the interval between the test and retest depends on the time needed to stop remembering details of the "test" (e.g., the answers given to specific test items), and varies in practice from minutes to (sometimes) years (e.g., in the case of stable personality traits).

Estimation of the test-retest reliability coefficient requires that the research units are ordered on the basis of their scores on the test and the retest; then the association among both orders is established (e.g., in terms of the Pearson correlation coefficient). This procedure draws heavily on the assumption that the structure of the concept under consideration does not change between the test and the retest. That is, the within-wave pattern of associations among the items should be the same across time, and it should still be plausible that these items all tap the same underlying concept. If this assumption of *structural invariance* is not warranted, it cannot be maintained that it is the same concept that is compared across time, and estimation of the test-retest stability is meaningless. The next section therefore discusses a conceptual framework to examine the assumption of structural invariance in longitudinal research.

3 A framework for examining structural stability in longitudinal research

3.1 Alpha, beta and gamma change

As suggested above, examining test-retest reliability presumes that the structure of the measures to be correlated is the same across time. The issue of across-time stability is often addressed in terms of a set of concepts coined by Robert C. Golembiewski and his colleagues (1976). They distinguished among three types of change, the first of which is *alpha change*, referring to ". . . a variation in the level of some existential state, given a constantly calibrated measuring instrument related to a constant conceptual domain" (p. 134). Alpha change (which should not be confused with Lee Cronbach's alpha coefficient for reliability) can be measured in terms of the degree of change across occasions, e.g., regarding the average height in a group of teenagers or the scores on two consecutive administrations of a mental ability test. Absence of alpha change implies that the test-retest reliability will be high; but, as said earlier on, this quantity is only meaningful insofar as the structure of the concept of interest has remained unchanged. This is what is meant by the reference to "a *constantly* calibrated measurement instrument" (italics added): the structure of the instrument should be invariant across time. For instance, if you step on a balance you are probably interested in knowing whether you had gained or lost weight, relative to the previous occasion when you measured your weight. The measurement of change occurs within a fixed system of stable dimensions of reality (the meaning of the concept of "weight" does not change), as defined by an indicator whose intervals are more or less constant (the calibrated marks on the scale of the balance).

Now imagine that the intervals between the marks on the scale of the balance would change across time. Then it would clearly be impossible to know whether you had gained or lost weight. Robert Golembiewski and his colleagues (1976) refer to such a phenomenon as *beta change,* defined as ". . . a variation in the level of some existential state, complicated by the fact that some intervals of the measurement continuum associated with a conceptual domain have been recalibrated" (p. 134). If beta change has occurred, chances are that the order of participants on the phenomenon of interest has changed, leading to low test-retest reliability. It is also possible that beta change has affected the pattern of associations among the items of your instrument, meaning that the internal consistency of your measure has changed across time. Of course, beta change regarding the scale of a balance is unlikely, but the scales of the balances used in social, educational and behavioral sciences (our items and scales) can certainly be subject to beta change. For instance, you may judge the characteristics of your house (e.g., the size of the rooms, quality of the neighborhood) differently after having lived in it for two days than after two years; a couple in marital therapy may judge each other differently after a session or four, even if nothing has changed. Instances of beta change in the social sciences often involve a change in the perspective of the study participants are involved in; given a clearer (or just a different) perception of the situation they may highlight different aspects of this reality, due to experience or maturational processes. These are common processes in longitudinal research, meaning that there is often the potential for the occurrence of beta change.

Finally, Robert Golembiewski and his colleagues (1976) referred to *gamma change* as involving "... a redefinition or reconceptualization of some domain, a major change in the perspective of frame of reference within which phenomena are perceived and classified, in what is taken to be relevant in some slice of reality" (p. 135). Whereas beta change refers to change in the intervals measuring a relatively stable dimension of reality, gamma change is a *quantum shift* in the conceptualization of this reality, manifesting itself in changes in the patterns of relationships among the components (items) of the measurement instrument (e.g., the number of dimensions of the measurement instrument). Clearly, across-time comparisons of scores on a measurement instrument (including the estimation of a test-retest reliability coefficient) are meaningless in the presence of gamma change.

3.2 Examining alpha, beta and gamma change

Golembiewski et al. (1976) not only coined the terms alpha, beta and gamma change, they also proposed to examine these types of change by means of factor analysis. Factor analysis refers to a variety of statistical techniques whose objective is to represent a set of observed variables ("items") in terms of a smaller number of unobserved (or "latent") underlying variables (also called *dimensions* or *factors*). Factor analysis can help researchers decide whether a particular set of items taps the same factor by providing information about the number of factors that can reasonably be assumed to account for the associations among the items, which variables belong to a particular factor, and how strongly the variables are affected by (or "load on") this factor. Generally speaking, a single factor will emerge if the associations among the items are about all equally high. Conversely, if there are multiple sets of items present that are highly intercorrelated among each other but not with items belonging to other sets of items, more than one factor will account for the data. If a single factor emerges, low intercorrelations among the items will translate into low factor loadings and a low reliability; the error variance on these items is large, relative to the variance they share (this shared variance is presumed to be indicative of the true score variance). Conversely, high intercorrelations lead to high loadings and a high internal consistency.

In examining alpha, beta and gamma change, the recent literature has almost exclusively relied on one particular type of factor analysis, namely *confirmatory factor analysis* (CFA) as implemented in programs such as LISREL, EQS and AMOS. Although these programs were initially developed for a very specific type of analysis (i.e., confirming one's *a priori* notions about the data using rigid statistical testing) they can also be very helpful in examining beta and gamma change; alpha change is usually tested using analysis of variance, although this could in principle also be achieved using the programs mentioned above. The two important virtues of CFA for examining beta and gamma change are that factor models can be specified on a high level of detail, and that statistical tests are available to test whether one factor model fits the data better than a competing factor model.

As regards the latter, programs for conducting CFA all provide a range of statistics that can be used to determine whether a particular model fits the data acceptably well. The best-known of these is the chi-square test, showing to which degree the "observed" variance-covariance matrix for the items resembles the "expected" variance-covariance matrix, i.e., the matrix that is expected on the basis of the fitted model parameters. Large (statistically significant) differences between these two matrices indicate that the model to be tested does a poor job in reproducing the observed item variances and covariances; conversely, small differences suggest that the fitted model may well closely resemble the unknown process that generated the data. Further, these tests can be used to compare various factor models, e.g., a one-factor model to a two-factor model or, more relevant to the study of beta and gamma change, whether the same factor model applies across all time points in the study.

CFA, gamma and beta change

As noted earlier, current research on beta and gamma change is strongly rooted in the confirmatory factor-analytic tradition. In this tradition, gamma change is examined in terms of across-time differences in the number of factors accounting for the data and in the pattern of factor loadings. Beta change is measured as change in the magnitude of the loadings of the items on the factors and/or change in the variances and covariances of the latent variables and observed items. Using CFA, researchers can assess which model accounts best for the data at each respective occasion. If the same basic model applies (i.e., the number of factors and sets of items associated with these factors is the same for each occasion), parts of the model (e.g., the factor loadings) can be constrained to be equal for all occasions. Comparison of the fit of the constrained model to that of the unconstrained model then reveals whether the imposed constraint is empirically plausible (i.e., whether the constrained part of the model is invariant across time; if so, no gamma and/or beta change has occurred) or not (meaning that gamma and/or beta change has occurred). Below I present and illustrate a simple three-step procedure to examine the invariance of factor structures across time using data from a two-wave study. The procedure can easily be generalized to multiwave studies. Similar procedures have been proposed by Kenneth Bollen (1989). Finally, note that Adam W. Meade, Gary J. Lautenschlager and Janet E. Hecht (2005) address the use of modern item-response theory in establishing structural invariance.

3.3 A three-step procedure to examine structural invariance

Basically, the aim of the first step to be taken is to see whether the variance-covariance matrix for the variables (items) of interest is equal for both time occasions. Across-time comparison of the *correlation* matrix would be inappropriate, as the variances of the items in the correlation matrix are standardized to equal 1.00; therefore, the chances of finding across-time differences in item and/or factor variables (i.e., beta change) would be minimized when the correlation matrix would be analyzed. A significant difference between the variance-covariance matrices indicates that (1) some form of gamma change has taken place; the number of dimensions or the pattern of factor loadings may have changed across time; (2) some form of beta change has occurred; the variances and/or covariances among the latent factors and/or the items have changed; or (3) a combination of (1) and (2) has occurred. The analysis would then proceed with the second step, examining what the precise differences across the occasions are; is the basic factor structure the same? If so, are the factor loadings the same? And so on. Conversely, if the test statistics suggest that there are no statistically significant differences among the variance-covariance matrices for both occasions, the associations among the items generalize across time, meaning that the factor structure has remained basically the same.

If the first step revealed that there are across-time differences in the factor structure, we must check in the second step whether the same basic factor structure applies to both occasions (in terms of the number of factors and the patterns of factor loadings). At both occasions a *simple structure* should have been reached, i.e., the simplest model (with the fewest factors) that still accounts reasonably well for the associations among the items should have been obtained. If the number of factors in the simple structures is the same for both occasions, we must examine whether the pattern of factor loadings is the same across time; for structural equivalence to hold, the same set of items must load at the same latent factor at both occasions.

If the same basic factor structure applies to both occasions, the issue of beta change can be examined in the third step. To test whether beta change has occurred, the loadings of the items on the factors must be constrained to be equal across time. Comparison of the fit indexes for the constrained and the unconstrained model will then reveal which of these models fits the data best. If the assumption of equality of factor loadings across time can be retained, the equality of the variances and covariances among the latent factors (the latter only if we have more than a single factor) across time can be examined, again by comparing a constrained model with an unconstrained model. This test focuses on the extent to which participants see greater integration or differentiation of constructs from one occasion to another. Finally, the equality of the error variances of the items can be tested.

3.4 Illustration: the structure of newcomer role ambiguity

As an illustration of the issues addressed above, we present a small example in which we probe the structural invariance of a three-item measure of role ambiguity among newcomers in their first job. These newcomers were interviewed twice; once after they had been in their jobs for on average six months (Time 1), and the second time after on average 30 months after entering this job (Time 2). Participants were recruited in eight European countries, total *N* at Time 1 was 2643; of these, 1245 participants (47.1% response) also participated in the second round of data collection. At both occasions, all participants completed a structured questionnaire in their respective languages, addressing work attitudes, work characteristics, and well-being.

One of the scales in this questionnaire tapped the degree to which the participants experienced *role ambiguity,* referring to the degree to which they were uncertain about several matters relating to their jobs (Table 9.1 presents the scale items). It is certainly not impossible

Table 9.1 Correlations and standard deviations (on the diagonal) of three indicators of role conflict among newcomers after being six months (Time 1) and 30 months (Time 2) in their first job (N = 1245)

	(1)	(2)	(3)	(4)	(5)	(6)
Time 1						
(1) …knows exactly what is expected [a]	1.11					
(2) …knows what s/he has to do [b]	.50	1.17				
(3) …procedures for handling things [c]	.42	.43	1.14			
Time 2						
(4) …knows exactly what is expected [a]	.30	.20	.17	1.17		
(5) …knows what s/he has to do	.20	.27	.15	.64	1.20	
(6) …procedures for handling things [c]	.17	.12	.23	.53	.52	1.15

Note: All correlations significant at $p < .05$.
[a] The full item was "On my job, I know exactly what is expected from me".
[b] The full item was "Most of the time I know what I have to do on my job".
[c] The full item was "On my job there are procedures for handling everything that comes up".

that the meaning of this concept changes across time. After six months in their first job, newcomers may still be in the process of finding out what is actually expected from them and how things should be done. In contrast, after 30 months they are reasonably experienced, and it would seem likely that in that stage of their careers they know perfectly well what to do, how that should be done and how possibly conflicting demands can be managed. This process of worker maturation may well reflect itself in the associations among the scale items. Table 9.1 presents the means, correlations, and standard deviations for the items of our role ambiguity scale.

The within-wave reliability coefficients (Cronbach's alpha) for these three items were .71 at Time 1 and .80 at Time 2, respectively. Although these alphas suggest that this three-item scale is reasonably reliable at both time points, it does not follow that it is the *same* phenomenon that is measured reliably by these items; factor loadings, factor and/or item variances may differ. This impression was confirmed through a comparison of the Time 1–Time 2 variance–covariance matrices. A test of the hypothesis that the elements of these matrices are the same across time yielded a chi-square value of 37.88 with 6 *df*, $p < .05$, thus rejecting this hypothesis: there are statistically significant differences between the variance-covariance matrices obtained and Time 1 and Time 2, meaning that there is reason to assume that gamma and/or beta change has occurred.

In the first step we examined whether a single-factor model for the associations among the three items applied to both occasions (model M1 in Table 9.2). In this model, all three items are presumed to load on a single underlying latent factor; this basic factor structure is expected to apply to both study waves. No constraints concerning the magnitude of the factor loadings, factor variances or item error variances are imposed. Table 9.2 shows that this model fits the data acceptably well, as evidenced by a non-normed fit index (NNFI) of .91. Thus, it seems that a one-factor model applies to both occasions; there is no evidence for the presence of gamma change.

Table 9.2 Analysis of invariance across time of role ambiguity among newcomers (N =1245)

Model		χ^{2a}	*df*	*NNFI*[b]
M1	Unconstrained model	108.03	8	.91
M2	Factor loadings constrained across time	109.39	10	.93
M3	Factor variances constrained across time	126.40	11	.94
M4	Item error variances constrained across time	146.40	14	.93

[a] All chi-square values significant at $p < .05$.
[b] Non-Normed Fit Index (NNFI): values of .90 and over signify acceptable fit.

In the second step we tested the across-time stability of the factor model. Model M2 examines whether the magnitude of the factor loadings is equal across time by constraining these factor loadings to be equal for both occasions. Comparison of the fit of M1 to that of M2 shows that the latter model results in a slightly higher chi-square value, whereas it has also more *df*s. The gain in degrees of freedom compensates for the loss of chi-square points; Chi-square$_{M2}$-Chi-square$_{M1}$ = 1.36 with 2 *df*s extra, $p > .40$. Thus, the model in which the factor loadings are constrained to be equal across time (M2) fits the data not significantly worse than the unconstrained model M1, meaning that at both occasions the three items load equally high on the latent factor. This is also reflected in the value for the NNFI that is higher for M2 than for M1.

Similarly, we can test whether the variance of the latent factor is the same across time. To this aim, model M3 (in which the factor variances are constrained to be equal across time) is compared against that of M2. In this case we find an increase of 17.01 chi-square points with a gain of only 1 *df*; thus, M3 fits the data worse than M2, $p < .01$. Further inspection shows that the factor variance increased somewhat across time. However, the NNFI for M3 is slightly higher than that for M2, suggesting that there are no major substantive implications of the significant increase in chi-square points; as our sample is fairly large, even small and uninteresting differences between the Time 1–Time 2 variance–covariance matrices are statistically significant. Thus, whereas there are some indications that the factor variance for role ambiguity increased across time, for the time being we assume that the factor model is essentially the same.

In the final step, we examined whether the item error variances were the same across time. To this aim, model M4 (in which the item error variances are constrained to be equal across time) is compared against that of M3. This results in a loss of 20.00 chi-square points with a gain of only 3 *df*s, $p < .01$. Moreover, the value of the NNFI decreases as compared to that of M3, suggesting that this statistically significant increase in chi-square points has substantive implications as well. Further inspection of the data shows that for all three items, the item error variances increased across time (which is also reflected in the standard deviations reported in Table 9.1).

These results indicate that there is some evidence of the presence of beta change across time, in that the dispersion of our participants increases across time. However, the structure of the measurement model (basic factor structure and factor loadings) does not change across time, meaning that participant scores on role ambiguity can safely be compared across time. The fact that Cronbach's alpha increased across time is due to higher intercorrelations among the items at Time 2; it would seem possible that this is caused by the increase of the item variances across time, as the factor loadings are essentially the same for both measurement occasions. Alternatively, it is possible that correlated error variances due to uncontrolled third factors account for the higher Time 2 correlations in Table 9.1. However, it is important to note that our results indicate that after partialling out these error variances (as we did in our confirmatory factor analyses), the reliability of the role ambiguity concept remains the same across time.

4 Regression to the mean

One interesting phenomenon that is intertwined with the reliability of measures in a longitudinal context is the *regression to the mean* or the *regression effect*. This effect was already noted in 1924 by the famous statistician Edward L. Thorndike. Beyond a certain medium range of initial values, responses tend to be the reverse of what one would expect: extremely low initial scores will be followed by an increase at a follow-up study, where extremely high initial scores tend to be followed by lower follow-up scores. Regression toward the mean is often discussed in terms of measurement unreliability. The score on a particular measuring instrument may not exclusively reflect a participant's true score on the underlying concept but other (incidental) factors as well; it could be that extremely high (or low) initial values are due to these incidental factors. If these incidental factors are absent during the follow-up measurement, participants' scores are likely to regress toward the actual "true" score on the concept of interest.

Example: Role ambiguity among newcomers

As an illustration, Figure 9.1 presents the Time 1–Time 2 scores of the newcomers discussed in Section 3 on role ambiguity. For simplicity, all scores were rounded to the nearest integer. If regression to the mean applies, participants

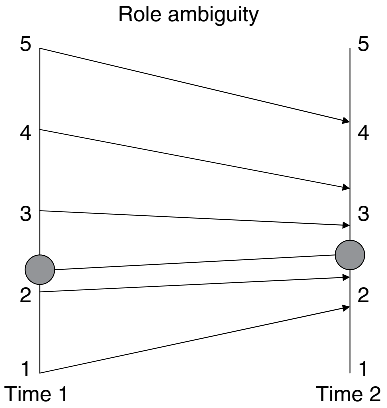


Figure 9.1 Regression to the mean among a longitudinal sample of newcomers, with respect to role ambiguity

having relatively high (low) scores at Time 1 should have lower (higher) scores at Time 2. The two gray dots in Figure 9.1 present the Time 1 and Time 2 means, respectively. As high scores represent low levels of role ambiguity, Figure 9.1 shows that the participants on average became less uncertain about what they had to do at their jobs. This general tendency applies strongly to those who felt extremely uncertain about their role (i.e., those with score 1) at Time 1; their scores are on average almost a full point higher at Time 2 (from 1.00 to 1.98).

One could easily argue that especially the group with low Time 1 scores on role ambiguity would be likely to become less uncertain about their roles. For example, these participants may ask their fellow workers or supervisor for clarification on their position or tasks. However, it is more difficult to understand why participants who were initially reasonably clear about their tasks would be so much more uncertain about their tasks at Time 2 (from 5.00 at Time 1 to 3.54 at Time 2). Regression to the mean would be a much simpler explanation for these findings than any substantive interpretation.

Example: Stability of rookie performances

Although regression to the mean is commonly considered a statistical artifact that distorts findings in longitudinal research, this effect may also be of substantive interest. Jim Taylor and Kenneth L. Cuave (1995) collected archival data on the performance of major league baseball hitters who had outstanding rookie seasons. If regression to the mean would apply (meaning that this excellent performance would be due to chance), outstanding rookies should have considerably less good second seasons. Taylor and Cuave (1995) found that the batting average of outstanding rookies declined from on average .300 in the first year to .276 in the sophomore year to .269 in the third to fifth year of their careers. Thus, outstanding rookie performances seem often the result of temporary factors that favorably influence their performance, meaning that the batting average is not a very reliable measure of the true qualities of rookie baseball players.

Dealing with regression to the mean

One possible strategy to deal with regression to the mean effects is to exclude participants with extreme scores on the initial measure of the phenomenon of interest; these are most likely to have their scores distorted by incidental factors. For example, in the newcomer example in Figure 9.1, one would exclude participants with Time 1 scores of 1 and 5. The intermediate scores (2–4) are probably less strongly influenced by regression to the mean than these extreme scores (although Figure 9.1 shows that participants with score 4 at Time 1 also experience a major decrease in the degree to which they are certain about their roles). Alternatively, one may increase the reliability of one's measures (by adding items or replacing bad items—note that this strategy cannot be applied during an ongoing survey, as this would endanger the across-time structural equivalence of one's measures).

5 Unreliability of change scores and the regression fallacy

In longitudinal research we are usually concerned about measuring and predicting the amount of change in the concepts that interest us (i.e., the prediction and description of alpha change, in the terminology used in the previous section). One natural way of measuring across-time change is the difference between the scores obtained at two time points. For example, in intervention studies researches are interested in the effects of a particular intervention. Clinical psychologists, for instance, may attempt to reduce feelings of burnout among teachers by training their time-management skills; comparison of the pre-intervention with the post-intervention scores (preferably accompanied with a comparison with the scores of a control group) then reveals whether the intervention was effective. Similarly, if the concept of interest is income, then subtracting the income measured at Time 1 from the income measured at Time 2 represents the income gain (or loss) during the Time 1–Time 2 interval. This quantity is termed the *difference score, gain score* or *change score*. Intuitively attractive as the change score may seem, its use has generated much concern among statisticians and methodologists. Prime among these concerns is the notorious *unreliability of change scores*; a related problem is known as the *regression fallacy.*

The unreliability of change scores

Statisticians have often attempted to discourage researchers from using change scores. Their critique has most eloquently been put forward by Lee J. Cronbach and Lita Furby in their influential 1973 paper, concluding that “there is no need to use measures of change as dependent variables and no virtue in using them” (p. 18). So what is the problem exactly? Assume that the reliabilities and variances of two repeated measurements of variable Y (Y_1 and Y_2, respectively) are the same for both measurements. The reliability of the Y_2–Y_1 change score is then given by

$$\frac{\rho_Y^2 - \rho_{12}}{1 - \rho_{12}}$$

with ρ_{12} denoting the correlation between Y_1 and Y_2, and ρ_Y^2 their reliability. Now, if ρ_{12} is positive (as is usually the case in longitudinal research; scores on concepts tend to be quite stable across time), it can be shown that the reliability of the difference score is always lower than the reliability of Y_1 and Y_2. Indeed, the higher the intercorrelation between both measures of Y, the lower the reliability of the difference score becomes. This is illustrated in Table 9.3, showing that if the reliability of Y_1 and Y_2 is very low (e.g., .30), it is impossible to have a decent reliability for the Y_1–Y_2 change score (reliabilities are considered acceptable when they are .60 or better; for diagnostic purposes, values of at least .70 and preferably .90 are required). For moderate (.60) and high (.90) reliabilities, the correlation between Y_1 and Y_2 becomes relevant; acceptable reliability for the Y_1–Y_2 change score is much easier to obtain when the correlation between the repeated measures is low. Unfortunately, the correlations among subsequent measures of the same concept tend to be high, usually being in the range

Table 9.3 Reliability of the change score as a function of the reliability of and intercorrelation among two repeated measures

Reliability of Y_1 and Y_2	*Correlation between Y_1 and Y_2*				
	.90	.60	.30	.10	.00
.90	.00	.50	.86	.89	.90
.60	-	.00	.43	.55	.60
.30	-	-	.00	.22	.30

of .50–.60 for a one-year interval (depending, of course, on the temporal stability of the concept under study; e.g., personality traits tend to be much more stable than moods). Thus, we often need unrealistically high reliabilities for the constituent variables to have a reasonably reliable change score. In the absence of such high reliability, the participants' scores on this change score may be little more than random error.

Of course, no one should be surprised by the typically low reliability of difference scores. The correlation between the Time 1 and Time 2 true scores on Y (i.e., the correlation between Y_1 and Y_2 after correction for the measurement error) is high for most regions of Table 9.3, as can be seen after application of the disattenuation formula $\xi_{12} = \rho_{12} / \rho_Y^2$. In particular, the true score correlation ξ_{12} equals 1.00 along the diagonal of zero reliability for the difference score. Clearly, the difference score approach cannot be expected to detect any true score-change in the absence of such change. This reasoning has fueled two approaches to increasing the reliability of the change score. First, Lee Cronbach (1984) argued that in order to measure the often small across-time true score change reliably, researchers should increase the reliability of their measures (e.g., by replacing bad items—that correlate only weakly with the other items of this measure—with better items or by increasing the number of items for their measures). This means that researchers will end up in the upper regions of Table 9.3, where it is easier to obtain an acceptable reliability for the difference score. Second, Ronald C. Kessler (1977) proposed that researchers make sure that the interval between the study waves is sufficiently large for true-score change to occur. This interval should be such that the difference between two consecutive measurements reflects at least partly true change and not just random fluctuations. Also note that increasing the interval between two measures of a variable leads to a lower correlation between these, meaning that researchers end up in the right half of Table 9.3, where it is easier to obtain acceptable reliabilities for the difference score. Unfortunately, the practical applicability of this advice is rather limited, in that it is usually unknown what the best length for this interval is. That is, a too-long interval may mean that participants' scores may change several times within this interval, making it difficult to relate the Time 1–Time 2 change to a predictor variable.

The combination of difference scores and the regression to the mean effect discussed in Section 4 can yield quite misleading findings. Kent M. Jennings and Gregory B. Markus (1977) were interested in the effect of having served in the army on feelings of trust towards the government. One possible strategy would have been to compare army veterans' trust scores to those of people without army experience. However, this simple design will not do as both groups may initially differ as regards their trust in the government. It seems likely that veterans put more faith in the government than nonveterans, or else they would not have chosen to join the army (the participants in Jennings and Markus' study had enlisted voluntarily). Thus, initial differences in trust must be controlled.

To this aim, Jennings and Markus conducted a two-wave longitudinal study, with the first wave being conducted in 1965 among high school seniors and the second in 1973. One way to analyze these data is to compare the Time 1 trust scores with the Time 2 scores, i.e., to compute change scores. However, as the amount of change tends to be negatively related to initial scores due to regression to the mean, this approach is fraught with difficulties. If veterans had higher Time 1 trust scores than others, the former group will presumably on average show smaller gains than the latter group on the null hypothesis of no effect. We would therefore conclude that serving in the army has a deleterious effect on trust in the government when, in truth, the null hypothesis is correct. This has been termed the *regression fallacy*, attributing the change in the criterion variables (that is presumably largely due to measurement unreliability) to the effects of the independent variables in the study.

In this case, one alternative for analyzing the data is the *regressor variable approach*. Instead of trying to relate the Time 1–Time 2 difference in Y to the scores on a predictor X, the Time 2 measure in Y is regressed on its Time 1 measure and X. Thus, the score on Y_1 is treated here like any other explanatory variable. In this vein, researchers can relate the change in Y to other variables (note that controlling for Y_1 will partial out what is constant in Y_2, thus leaving what has changed in Y to be accounted for by the other predictor variables).

Change scores: Present and future

At present, it appears that since the middle of the 1970s psychometrists' negative attitude towards using change scores has tempered somewhat. According to K.K. Sharma and J.K. Gupta (1986), difference scores can be quite reliable under commonly present circumstances, while Scott E. Maxwell and George S. Howard (1981) found that they may yield powerful tests of causal hypotheses, in spite of their unreliability. Similarly, Kim May and James B. Hittner (2003) argued that change scores yield powerful significance tests in that they reduce "true score" variance. Yet, although the difference score has been rehabilitated somewhat (see, for example, the discussions in Greenberg, Chapter 17, Twisk, Chapter 18, and Finkel, Chapter 29 in this volume), many cautious researchers will refrain from using change scores, if only because their readers (and especially reviewers) of their work may still be suspicious of this approach. In this sense, it may take a long time before we will witness renewed interest in using change scores in applied research.

6 Concluding remarks

The present chapter dealt with several issues relating to the reliability of measures that are applied in the context of two- or multiwave studies. As in single-wave (cross-sectional) research, it is imperative that the measures that are used have an acceptable reliability. If such is not the case, associations between variable pairs will be underestimated, leading to often disappointing null findings that are actually due to imperfect measurement. Apart from this problem (that applies equally strongly to cross-sectional and longitudinal research), we have shown that unreliability poses even more threats to the validity of findings in longitudinal research. In that sense, it is often sensible to invest much time and effort in selecting reliable measures to be included in longitudinal research—perhaps even more than one would for a cross-sectional study.

References

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Cronbach, L. J. (1984). *Essentials of Psychological Testing*, 4th edn. New York: Harper and Row.

Cronbach, L. J. and Furby, L. (1973). How we should measure change—or should we? *Psychology Bulletin,* 74: 32–49.

Golemiewski, R. T., Billingsley, K. and Yeager, S. (1976). Measuring change and persistency in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioural Science*, 12: 133–157.

Guion, R. M. (2002). Validity and reliability. In S. G. Rogelberg (ed.), *Handbook of Research Methods in Industrial and Organizational Psychology*, pp. 57–76. Malden (MA): Blackwell.

Jennings, M. K. and Markus, G. B. (1977). The effect of military service on political attitude: A panel study. *American Political Science Review*, 71: 131–147.

Kessler, R. C. (1977). Use of change scores as criteria in longitudinal survey research. *Quality and Quantity*, 11: 43–66.

Maxwell, S. W. and Howard, G. S. (1981). Change scores: Necessarily an anathema? *Educational and Psychological Measurement*, 41: 747–756.

May, K. and Hittner, J. B. (2003). On the relation between power and reliability of difference scores. *Perceptual and Motor Skills*, 97: 905–908.

Meade, A. W., Lautenschlager, G.J. and Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, 5: 279–300.

Nunnally, J. C. (1978). *Psychometric Theory*, 2nd edn. New York: McGraw-Hill.

Sharma, K. K. and Gupta, J. K. (1986). Optimum reliability of gain scores. *Journal of Experimental Education*, 54: 105–108.

Taylor, J. and Cuave, K. L. (1995). The sophomore slump among professional baseball players: Real or imagined? *International Journal of Sport Psychology*, 25: 230–239.

Thorndike, E. L. (1924). The influence of chance imperfections of measures upon the relationship of initial score to gain or loss. *Journal of Experimental Psychology*, 7: 225–232.

This page intentionally left blank

Chapter 10

Orderly change in a stable world: The antisocial trait as a chimera

Gerald R. Patterson

1 Introduction

A developmental perspective implies that changes in social behavior are related to age in an orderly way. In the present report, the definition is expanded to include any changes in social behavior accompanied by explanatory variables that account for significant amounts of variance in the change score. The experiences that bring about change may or may not be related to the age of the child.

By definition, the availability of longitudinal data sets would be prerequisite to the study of change in social development. What gives the developmental perspective credibility is that highly sophisticated techniques for analyzing longitudinal data sets have been well understood for over a decade. For example, the lucid descriptions of time-series analysis and panel analysis for inter- and intraindividual longitudinal data sets were described by Nesselroade and Baltes (1979). Collins and Horn (1991) detailed further developments in the analysis of change, such as latent growth modeling (LGM), analysis of factor invariance, event-history analysis, and the Guttman longitudinal simplex. The analytic tools are there; in fact, they have been there for some time.

Given the long-standing commitment to a developmental perspective and 40 years worth of longitudinal data, one might expect that the construction of an empirically based theory for the development of children's social behaviors would be well under way. In fact, there is no such accumulative data base. In the area of delinquency, for example, Farrington (1986) identified 11 well-designed longitudinal projects. The most salient finding to emerge was that measures of children's antisocial behavior were significant predictors for adolescent delinquency (i.e., antisocial behavior is highly stable). Nevertheless, efforts to explain the stability or to predict change in antisocial behavior (desistors or late starters) have not been particularly successful (Farrington and Hawkins, 1991). With few notable exceptions, even the more recent studies assiduously avoid the study of change. This author's review of the empirical findings leads to the conclusion that the developmental emperor either has no clothes or, at the very least, is prone to indecent displays.

Reprinted from *Journal of Consulting and Clinical Psychology*, 61(6): 911–919, 1993

1.1 Beyond stability coefficients

The meager returns from longitudinal studies reflect the interactive contribution of three errors in research strategy. First, most investigators have been trained within a myopic conceptualization of the trait as a static or fixed unit that can be satisfactorily assessed by data from a single agent. In the discussion that follows, formulation of the trait concept is expanded to include changes over time. The measurement is based on multiagent, multimethod data. The second limiting factor is found in an overweening reliance upon the correlation coefficient as the analytic tool. Being wedded to both of these strategies leads unerringly to over-production of stability coefficients as the main output for longitudinal studies. The major limitation in strategy, however, lies in failure to include a formulation about the nature of change in social behaviors. This failure may be due to the fact that developmental theories are exceedingly vague about what produces change and how to measure it. Most formulations consist of ambiguous metaphors about organismic variables, social learning, family systems, or attachment. Carrying out developmental research under the aegis of these three limitations is analogous to studying a dance through a tube. Not knowing where to look, we arbitrarily collect periodic data on the kind of shoes being worn. In so doing, we not only miss the point of the process, but produce yet another set of very high-order stability coefficients saying that people tend to wear the same shoes throughout the dance.

In the alternative perspective, the present report uses multiple indicators to define the trait concept. This strategy reflects the growing consensus that any single indicator measure would be systematically biased (Patterson, 1992; Patterson and Bank, 1987; Sullivan, 1974). The contemporary shift toward using multiple indicators and confirmatory factor analyses (Bentler, 1980) supposedly provides a basis for constructing models that are more generalizable. The utility of this strategy has been demonstrated by replicated models from a cross-sample (Forgatch, 1991) and across-site (Conger, Patterson, and Gé, 1995) studies. From this perspective, a trait such as antisocial behavior is embedded in a matrix of changing social behaviors (Patterson, Reid, and Dishion, 1992). Not only is the matrix changing, but some forms of the trait itself are changing. A study of this process requires having a theory about what produces the trait, how it will change, and when. Patterson, Reid, and Dishion (1992) detailed an empirically based theory about boys' aggression. The theory specifies what mechanisms produce changes and, to some extent, when these changes should occur. These considerations led to an expanded definition of the trait score requiring time of emergence as a necessary piece of information. For example, an early emergence for the trait increases the risk for qualitatively new problems directly caused by the problem child's coercive and antisocial behaviors. Literature that examines the covariation between the trait measures and these qualitative changes is reviewed in a section that follows.

A core problem for a developmental theory of aggression is the need to explain the systematic changes in form and in intensity of these behaviors. Changes in form occur at all stages of development, but particularly during adolescence. In this report, LGM is used to examine the increment in growth for two antisocial behaviors (truancy and drug use), and covariates that explain why these changes for new forms come about are examined (Stoolmiller and Bank, in press).

In this interpretation, the child's antisocial trait is viewed as a *chimera*. According to biologists, a chimera is an unusual hybrid produced by grafting tissue from different organisms. This metaphor is an apt descriptor for the antisocial trait. Each addition of a qualitatively new problem, and each change in the form of the coercive or antisocial behavior, might be thought of as a graft made onto the original trait score. If changes do occur, how can we say we are describing the "same thing"? Analyses are presented demonstrating that qualitative additions and changes in form define a second-order deviancy factor changing in an orderly manner over time.

Before moving to a discussion of findings, one further problem in strategy needs to be raised. A frequent claim made for longitudinal design is that it can make an important contribution to evaluating causal status for developmental variables (Nesselroade and Baltes, 1979). Gollob and Reichardt (1987) extended this model to an autoregressive design where variable *x* measured at T1 is partialed out of the measure for *x* at T2. They asserted that a panel design could be used to test for the causal contribution of variable *y* measured at T1 by demonstrating that it covaried significantly with future changes in *x*. Cross-lag correlations could be used to test the hypothesis that *x* measured at T1 could demonstrate a causal effect on *y* measured at T2. Rogosa (1979) carefully delineated problems in measurement and specification that typically make this a very weak test for causal status. He also strongly endorsed multiple measures and the use of structural equation modeling (SEM) rather than traditional multiple-regression analyses. Testing the causal status of parenting practices was, in fact, one of the prime goals of the longitudinal Oregon Youth Study (OYS) (Patterson, 1988). The OYS was designed to fulfill the requirements Rogosa outlined. As we applied the autoregressive format suggested by Gollob and Reichardt (1987) to our longitudinal data, however, we were immediately confronted with a paradox. Using latent constructs to measure child traits and parenting practices routinely generated stability coefficients ranging from .70 to .85 for two- to four-year intervals (Patterson and Bank, 1989). Correlations with causal constructs were usually well above the .40 to .50 range. Paradoxically, our efforts to use auto-regressive panel models to test for causal status were doomed to failure because our measures were too good! Stoolmiller and Bank (in press) point out that the combination of high stabilities and high colinear relations make significant cross-lag effects extremely unlikely. They also point out that there are alternative analytic strategies which are more likely to be effective (e.g., LGM). The present author believes that, at best, SEM or LGM can provide only a weak test of causal status. A developmental theory must eventually be based on experimental evidence. For example, Dishion, Patterson, and Kavanagh (1992) and Forgatch (1991) have described longitudinal designs that included random assignment and experimental manipulations to test for causal status.

Longitudinal data from the OYS are used to address questions about quantitative and qualitative change. The 206 families involved in the OYS live in high-risk (for crime) neighborhoods in a medium-sized metropolitan area. The recruitment procedures and sample characteristics were described by Capaldi and Patterson (1987). Each family participated in over 20 hours of assessment at each probe, when the boys were in grades 4, 6, 8, and 10.

2 Stable but changing

When studying changes in children's antisocial behavior, we must first establish a small island of stability. We begin by examining two very different facets of stability. First, the stability of the definition itself is considered; it may be that what is meant by stability changes as the individual moves from childhood through adolescence. The second question concerns the means for estimating the magnitude of stability coefficients. There is some reason to believe that the typical bivariate correlation of monoagent reports might overestimate stability.

2.1 Stability in definition

It may be that both the form and the definition of antisocial behavior change as the child's age increases. Eddy, Heyman, and Weiss (1991) partialed out the effect of changes in form by constructing a pool of antisocial behaviors that could occur at any age. Maternal ratings of these 12 items were available for toddlers (age 27 months) and for five-year-olds. The items were rank-ordered at each age according to frequency of use. The correlation of .78 showed that items most frequently used to describe toddlers were also more likely to be used for preschool children. This same set of items, when scored for different samples of boys in grades 2 through 8, showed comparable stabilities in definition. For example, the rank ordering of items for toddlers correlated in the .71 to .75 range with the rank ordering of items used by mothers to describe their adolescent sons. In each case, the most frequently checked items tended to be "disobedience" and "temper tantrums," whereas the least frequently checked was "physical aggression." What mothers perceive to be "most" and "least" antisocial remains constant over child and adolescent development.

In the Eddy et al. (1991) article, mothers' stability in definitions for boys between grades 4 and 8 was .96. For that same sample, the individual difference stability for mothers' ratings of their sons was .65. By definition, however, the error terms for the two sets of ratings were intercorrelated. Not only does this violate a fundamental assumption for application of correlational analysis, but it also inflates the magnitude of the correlation. How does one partial out the joint contribution of shared method variance and stability? As Rogosa (1979) and others have pointed out, the use of trait indicators based on reports from multiple agents and methods would enable us to disentangle the contributions of shared method and stability variances. This possibility is examined in the following section.

2.2 Stable individual differences

The preceding analyses showed there is a core set of interpersonal reactions defining the antisocial trait that are stable over the period from age 10 through age 14. Adolescent boys, however, are involved in some antisocial acts that younger boys seldom engage in, so the definition was expanded to include all of the antisocial acts that are performed often by adolescents. SEM was used to estimate the stability of antisocial behavior for the 206 boys in the OYS. The trait was defined by parent reports, teacher reports, and child self-reports at grades 4 and 8. Details of the psychometric analyses from the grade 4 studies for each of the indicators and for the construct itself were presented by Capaldi and Patterson (1989).

In the model, when the same methods were used at the two assessment probes, the error terms were allowed to covary. In doing this, the contribution of shared method variance can be removed from the estimate of stability. As shown in Figure 10.1, all of the factor loadings were highly significant at both points in time; in fact, the loadings showed at least configural invariance. At both assessments, the highest loadings were for parent reports and the lowest were for child self-reports. The probability value of .63 for the chi-square test showed a solid fit between the data set and the a priori model.

The traditional equation for test-retest correlation $(1-r^2)$ defines the error of measurement for a given trait score. Although the current stability path coefficient suggests that the error of measurement might be about 31% of the variance, it is plausible that a significant proportion of this error estimate might reflect changes in level for a subgroup. A portion of the unexplained variance could reflect systematic changes in individual growth over time. In the next section I examine the possibility of systematic changes under conditions of high stability.

3 Two developmental models

A formulation about late-starting delinquents (Patterson, DeBaryshe, and Ramsey, 1989) suggested that there would be substantial numbers

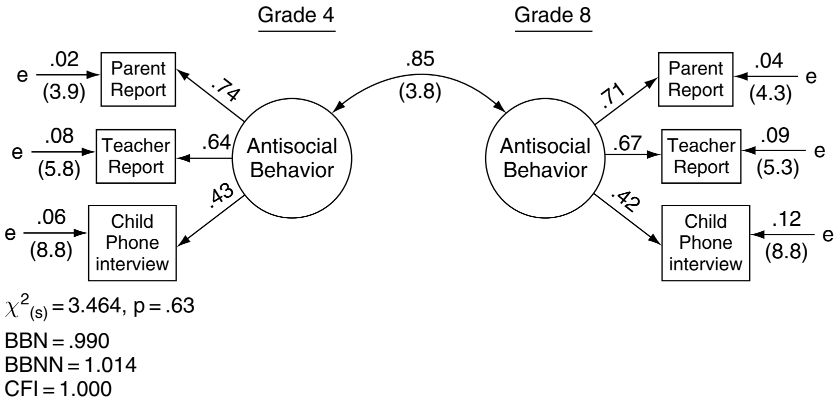


Figure 10.1 Stability of antisocial behavior from grade 4 to grade 8

who become antisocial for the first time during midadolescence. This would result in significant shifts in mean level for antisocial behaviors; presumably, the resulting shifts in ordinal rankings would in turn contribute to lowered stability coefficients. Intraindividual growth curve analyses will be used to (a) demonstrate that such systematic shifts do occur and (b) identify a set of variables thought to bring this about.

It was hypothesized that early- and late-onset antisocial boys represent two very different groups. Each group has different determinants and outcomes. A two-parameter latent growth model would seem useful in examining this general model. For example, one parameter (the intercept) would define where the process starts (i.e., the early onset boys). The other parameter would describe how intraindividual growth unfolds over time (i.e., the late-onset boys). In the present context, the intercept describes the individual differences in the antisocial trait at age 10. The second parameter constitutes an operational definition of late starter (i.e., nonproblem children who developed antisocial traits in early adolescence).

Part of the beauty of the latent growth model is that it operationally defines *early* and *late* starters. In addition, it introduces covariates that might account for the variance in the intercepts as well as the individual differences in growth (i.e., provides a correlation test for potential determinants). The covariates provide a direct test for the assumption that determinants for one developmental phase (early starters) may be different from the determinants for another phase (late starters). Initially, Patterson et al. (1989) hypothesized that determinants for early starters would be provided by contingencies embedded in family interaction; these in turn are controlled by the effectiveness of parenting practices. Latent constructs for parental effectiveness in discipline and monitoring practices assessed at age 10 serve as covariates accounting for individual differences in intercept scores for the OYS.

A second parameter in the LGM, the shape parameter, described the differences among the boys in intraindividual growth patterns for antisocial behavior assessed at grades 4, 6, 7, and 8. These late starters did not begin their antisocial careers until early adolescence; they showed few, if any, adjustment problems during childhood and possessed at least marginal social and survival skills (Patterson and Yoerger, 1993). The assumption is that some of the late starters will be arrested, but few will be chronic delinquents, and they will have a better prognosis than early starters for moderate levels of adult adjustment. The mechanisms that determine the delinquent behavior for late starters are an earlier than normal presence on the street (i.e., *wandering*) and a heavy commitment to the deviant peer group. This formulation was based on Stoolmiller's (1990) use of the OYS longitudinal data to demonstrate that, during the interval from childhood to midadolescence, changes in antisocial behavior covaried with changes in wandering and changes in involvement with deviant peers. The hypothesis about the involvement of deviant peers in direct training for delinquent acts is tested in a later section. Multiple indicators are used to define the constructs for wandering and involvement with deviant peers that would presumably covary with changes in individual growth for antisocial behavior (Rogosa and Willett, 1985).

3.1 Changes in antisocial behavior

For the entire OYS sample, teachers' ratings for antisocial behavior showed a significant increase in mean level from grade 4 to grade 5, but no significant change over the next four years (Patterson, 1992). Current studies using parent ratings and child self-report data showed the same general pattern (i.e., essentially a non-significant slope for measures from grade 4 through grade 8). The finding of no increase in antisocial behavior from early to midadolescence is consistent with the findings from teacher and adolescent ratings in the Chapel Hill longitudinal study (Cairns and Cairns, 1991).

In the OYS, the same assessment battery was used at grades 4, 6, 7, and 8. At each point in time, the raw scores from teacher, parent, and child telephone interviews were added to generate a single score. The same set of items was used at each point in time.

3.2 A two-factor latent growth model

The results from the simultaneous test of the early- and late-starter models are summarized in Figure 10.2. Using a two-factor model requires that the correlation between the slope and intercept parameters be minimal or zero. In fact, the data showed that the initial level for the antisocial score was unrelated to slope changes. This is a crucial piece of information for a developmental model of antisocial behavior; where a child starts is not necessarily related to his or her future growth (in mean level).

The error terms for each wave of measurement for the antisocial construct were set equal. The intercept was defined by the antisocial scores obtained at grade 4; the factor loadings for the antisocial measures were set at 1.0 for each wave. Based on prior work by Stoolmiller (1990), the factor loadings for the shape parameter were set at zero for grade 4 and at 2 for grades 6, 7, and 8. All the error terms for the covariates were allowed to covary.

The data showed that measures collected at the four points in time defined the same latent construct for antisocial behavior. Evidence for this assertion lies in the r^2 values (adjoining each measured variable), which ranged from .68 at grade 4 to .76 at grade 8. These values show that the latent construct loads at a very high level on the measures at each point in time.

It was hypothesized that there would be a significant contribution by the discipline and monitoring constructs in accounting for initial level of child antisocial behavior. The findings were consistent with the hypothesis. Ineffective parental discipline (−.44) was associated with higher intercept scores for antisocial behavior even after the contribution of inept monitoring had been partialed out. The comparable path coefficient for monitoring was −.37. Together, the parenting practices accounted for 35% of the variance in the initial level of antisocial scores. Neither set of initial parenting skills covaried with late intraindividual changes in antisocial behavior. The results suggest that a single-minded focus on parenting skills is helpful in understanding initial levels of aggression, but it does not tell us very much about which boys will be a risk for an increase in antisocial behavior at midadolescence.

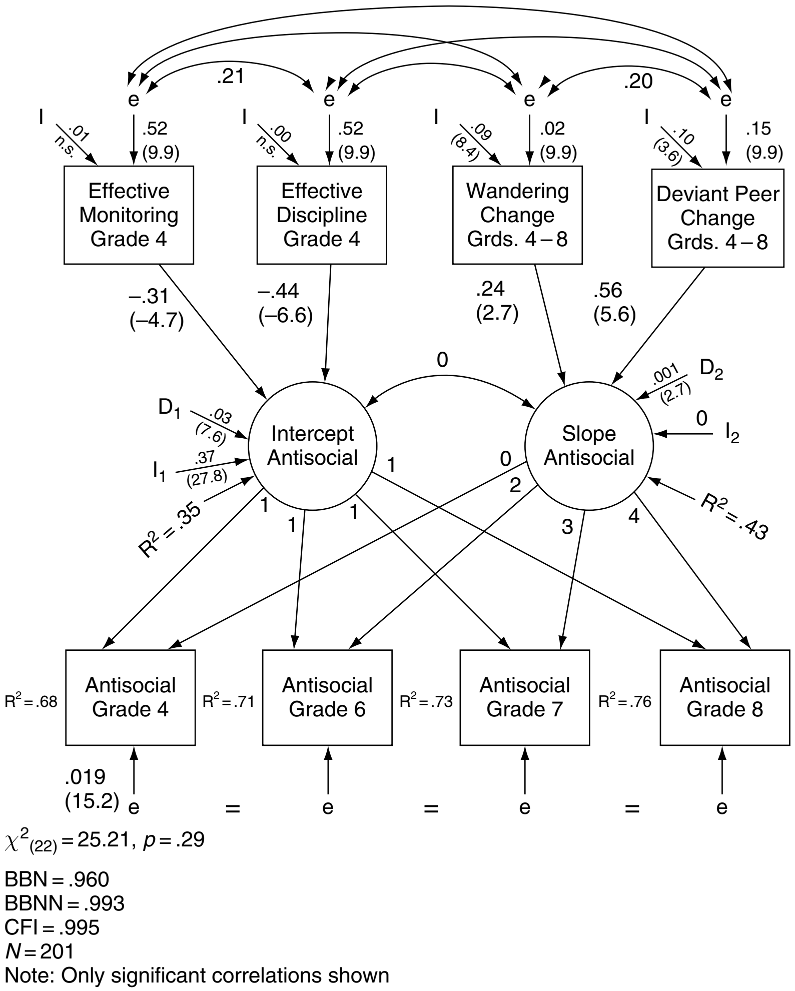


Figure 10.2 Developmental models for early and late starters

The next hypothesis examined in Figure 10.2 was that the boys who showed increases in wandering (i.e., unsupervised street time) and in involvement with deviant peers would also show increases in antisocial behavior. Changes in wandering and in involvement with deviant peers were expressed as a simple difference score (Stoolmiller and Patterson, 1995). The difference scores showed a mean increase in wandering of .09 from grade 4 to grade 8 and a mean increase in deviant peer involvement of .10. The increases in wandering contributed significantly (.24) to the slope index for increases in antisocial behavior. Increases in deviant peer involvement contributed heavily (.56) to changes in antisocial behavior even after the contribution of increased wandering had been partialed out. The combined contribution of the two variables accounted for 43% of the variance in the measures (shape parameter) of intraindividual growth in aggression.

The nonsignificant chi-square value demonstrated an acceptable fit of the data to the a priori early/late starter model presented in the original statement (Patterson et al., 1989). The implication is that there are two developmental paths for antisocial behavior. As hypothesized, each path is characterized by a different set of covariates. The path of the parenting skills model relates significantly to the start point, childhood aggression measured at age 10. The path of the deviant peer involvement model relates to antisocial behavior that begins in early adolescence.

Early and late starters seem to have different determinants. It was also hypothesized that the two groups would differ significantly in the timing for their first arrest and for their risk of chronic offending (Patterson et al., 1989). Presumably, antisocial children would be at greater risk than late starters for both early police arrest and for chronic arrests during adolescence. The longitudinal data collected for the OYS provided strong support for both hypotheses (Patterson, Crosby, and Vuchinich, 1992; Patterson and Yoerger, 1993).

The findings for the late starters' intraindividual growth in antisocial behavior suggest that some of the variance not accounted for in stability estimates may be generated by systematic increases in aggression for subgroups of young adolescents.

4 Analyses of qualitative shifts

It was hypothesized that over time there are two major qualitative changes in problem behaviors that accompany the antisocial trait. One shift is in the form of the antisocial acts; the other involves the addition of nonantisocial problem behaviors. The assumption being tested is that these qualitative shifts are quantifiable and that they define an emerging second-order deviancy factor.

4.1 The addition of new problem behaviors

Coercive and antisocial acts elicit predictable reactions from the social environment, adding qualitatively new problems to the developmental trajectory (Patterson et al., 1992). The additions occur in an identifiable sequence of reactions by members of the child's social environment. The sequence is initiated by the entrance of the antisocial child into the school setting. In that setting, the child's coercive interpersonal style produces an immediate reaction. Coie and Kupersmidt (1983) showed that, in a newly formed group, other children began to label the antisocial child as "disliked" within 2 or 3 hours of contact. The second reaction to the child's behavior is from the teacher. The child's obdurate noncompliance to implicit and explicit rules means she or he spends less time on tasks when in the classroom and less time on homework assignments. The child's academic failure is probably evident to him or her early on. By grades 3 or 4, the antisocial child has failed in two fundamentally important tasks: peer relations and academic skills. Patterson and Capaldi (1990) hypothesized that the effect of this dual failure is increased frequency of depressed moods. School failure, peer rejection, and depressed mood constitute a cascade of qualitative problems that add to the child's burden. The link between prior antisocial behavior and academic failure and peer rejection has been examined by SEM in a series of analyses summarized by Patterson and Yoerger (1993). The relation between dual failure and depressed mood has been replicated in three SEM studies detailed by Patterson and Capaldi (1990) and Patterson and Stoolmiller (1991).

4.2 Developmental changes in form of antisocial acts

The stability studies reviewed earlier imply that the antisocial acts of a five-year-old may be prototypic of the acts of the delinquent adolescent. It is evident, however, that there are profound changes over time in the form of coercive (e.g., noncompliance, threats, temper tantrums) and antisocial (e.g., stealing, lying) acts. How do we move from the five-year-old's noncompliance and temper tantrums to the adolescent's substance abuse, burglary, and shoplifting? We believe that many crucial changes in the form of antisocial acts occur during early adolescence, and the primary agents of change for these qualitative shifts are members of the deviant peer group. During early adolescence, the training is fairly intensive and may involve multiple antisocial acts. If this is so, several new forms of antisocial behavior might change in a similar fashion over time. To test this hypothesis, data were examined for the changes in substance use and in truancy at grades 4 through 9. Substance use was defined by a single item, "uses alcohol or drugs" (never, sometimes, very often), from the Child Behavior Checklist (CBCL) (Achenbach and Edelbrock, 1983) filled out by one or more teachers at each grade. The truancy variable was based on ratings for a single item, "skips school" (not true, sometimes true, often true), from the CBCL. The ratings were made by mothers in single-parent families and by both parents in intact families. For both variables, there was about a fivefold increase from grade 4 to grade 9; much of the growth occurred between grades 7 and 8. The key hypothesis is that the individuals who show growth in one form will also be at significant risk for growth in the other.

The results of the two-parameter LGM are summarized in Figure 10.3. The initial phases of the growth were characterized by a very high incidence of zero values. This violation of the assumption of normal distributed variables is a major cause for concern. Although the assumption of equal error variance could not be met, setting the error as proportional to variance proved successful. The probability value of .15 for the chi-square value of 34.59 (df = 27) showed an acceptable fit of the a priori model to the data set.

Only a few boys showed positive initial scores for the truancy and substance use constructs. As shown by the path coefficient of .34, there was a significant likelihood that these early starters were involved with deviant peers. The timing of the involvement with deviant peers seems to be critical, as evidenced by the path coefficient of .58 between early involvement (age 10) and later growth in the new forms of antisocial behavior: the earlier the onset, the greater the future growth. However, increasing involvement with deviant peers also contributed significantly to individual difference variance in growth. Note that the timing and the growth in deviant peer involvement made significant contributions of the same magnitude. Increased wandering also contributed significantly. Taken together, the information from the three covariates accounted for 54% of the variance in the slope factor.

4.3 Quantifying qualitative shifts

The behaviors that define the antisocial trait may serve as determinants for a host of new problems such as peer rejection, academic failure, and depressed mood. New forms (e.g., truancy, substance abuse, police arrest) are constantly being added. This raises the question of whether these changes themselves form an orderly pattern of change over time. One way of thinking about this problem is to consider each qualitative change as contributing to a second-order deviancy factor that has the antisocial trait as its core. As each qualitatively new problem emerges and the prevalence becomes noticeable, it should appear as a newly significant factor loading on a second-order deviancy factor. At younger ages, indicators for qualitative shifts such as substance abuse and arrest would describe so few members of the sample that the factor loadings would be nonsignificant. Repeated factor analyses at early and midadolescence should show that each qualitative shift eventually loads significantly on the second-order deviancy factor. Each addition is but a new branch of what is essentially the same thing.

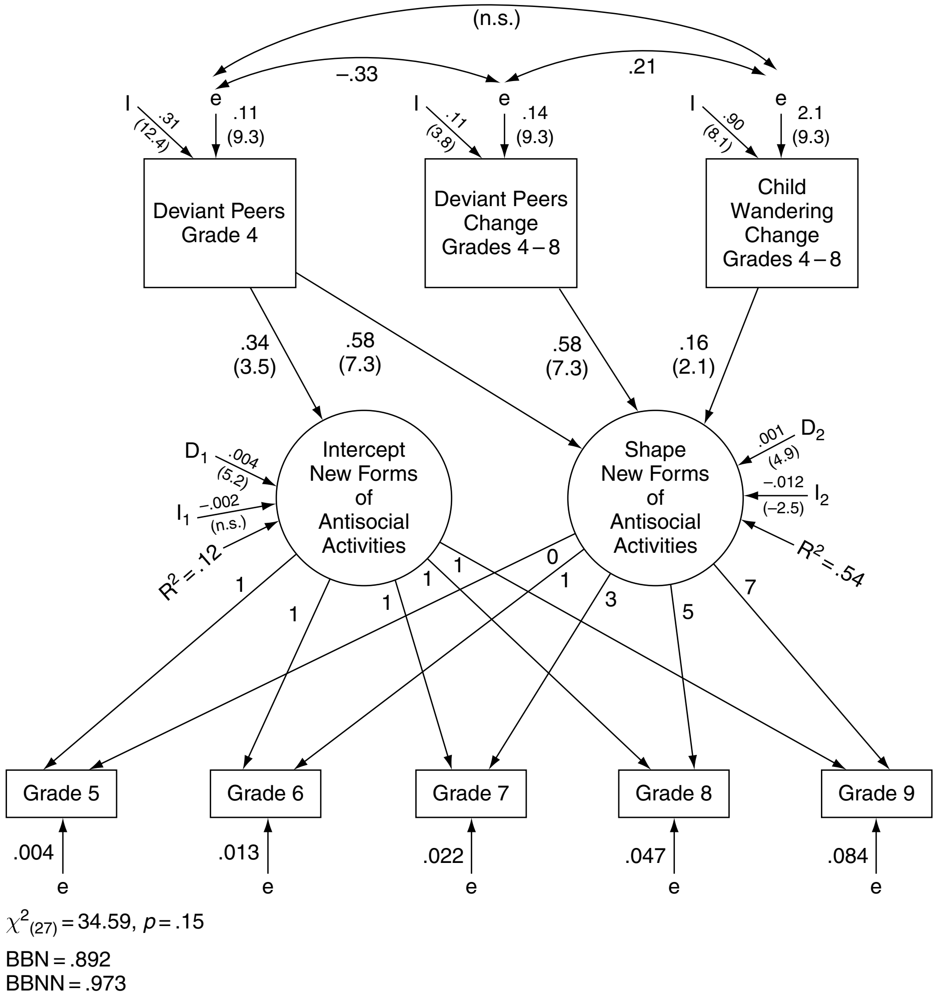


Figure 10.3 A growth model for changes in form

By way of illustration, four qualitative shift variables from the OYS were each measured at three points in time. Two of the variables—antisocial behavior and academic failure—were latent constructs; the other two—police arrests and teacher ratings of substance use—were new forms. The findings for the data collected at grades 4, 6, and 8 are summarized in Figure 10.4. At each point in time, the antisocial construct serves as the core defining variable; the estimated factor loadings for waves 1 through 3 were 1.0, 1.0, and .89 respectively. Over time, the academic skills construct makes an increasing contribution to the deviancy factor; the loadings shifted from .32 at the beginning of the process to a substantial .43, and then .60. The other variables representing qualitative shifts (substance use and police arrest) also showed increased loadings on the deviancy factor. The data sets at all three points in time

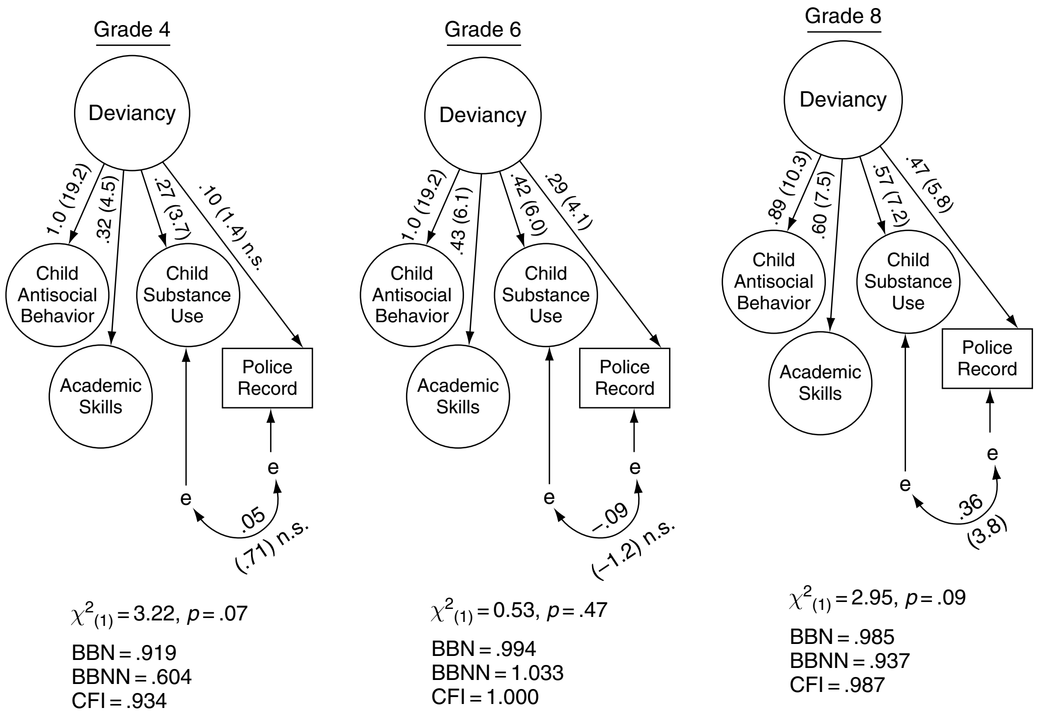


Figure 10.4 Changes in the structure of deviancy over time

provide a good fit to the model, as shown by the nonsignificant chi-square values.

The findings demonstrate that the factor structure defining deviancy is altered significantly between age 10 and age 17. However, the second-order factor analyses suggest that these qualitative changes represent a pattern of orderly change over time.

5 The trait as a chimera

The antisocial trait defines an interpersonal style that may maximize short-term gains but adds to long-term increases in misery (Patterson et al., 1992). An understanding of the trait requires both a short-term and a long-term perspective. The second-order deviancy factor is one means for expressing in quantitative terms the nature of some of these long-term changes and demonstrating that the changes are systematic.

The chimera metaphor implies that these qualitative changes are analogous to tissue grafts. The Greeks described the chimera as a fire-breathing creature that was part goat, part lion, and part snake. As shown in Figure 10.5, from a developmental perspective, the chimera begins as a goat and always retains its goat-like essence. This is analogous to seeing the antisocial trait as the underlying essence for the deviancy factor. Academic failure and peer rejection components constitute the addition of the lionesque countenance to what is essentially still a goat. By midadolescence, the additions of substance use and police arrest produce an aroused society and complete the conversion of a simple goat to a fire-breathing monster with the tail of a snake.

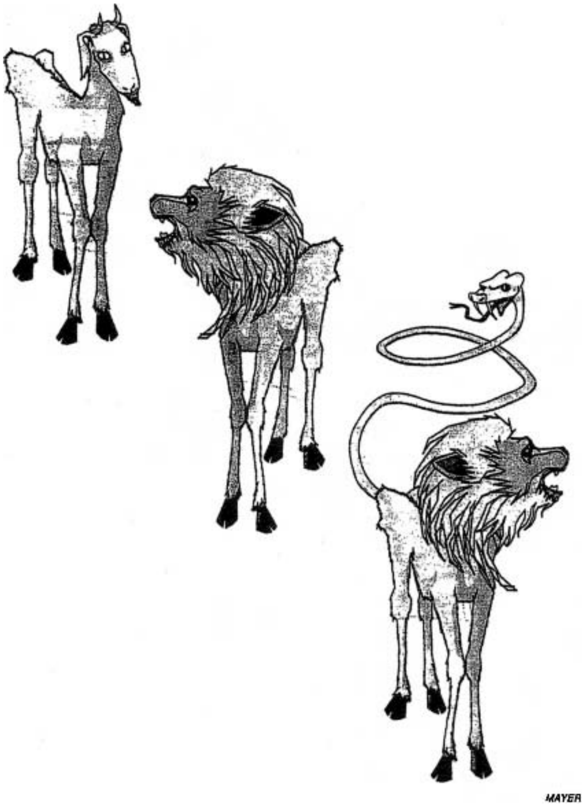


Figure 10.5 The chimera effect

6 Implications

In summary, a developmental approach to the study of antisocial behavior requires not only longitudinal data and a theory but also a sensitive application of statistical analyses that consider intraindividual growth over time. The concept of growth must be expanded to include quantitative changes in mean level over time as well as qualitative changes generated by the same process. A trait is generalizable across time and across settings but, in a very real sense, it reflects an underlying dynamic process.

Acknowledgments

I gratefully acknowledge the contribution of Grant No. MH 37940 from the Center for Studies of Antisocial and Violent Behavior, NIMH, US, PHS, in collecting and processing the longitudinal data sets for the OYS. I also wish to acknowledge the support from MH 46690 from the Prevention Research Branch, NIMH, US PHS, which made possible much of the analyses.

I am very grateful for the exchanges with my colleagues Karen Yoerger and Mike Stoolmiller. Without those exchanges, this manuscript simply would not have come about. I wish to thank Beverly Fagot and Lew Bank for their critiques of an earlier version of the manuscript.

References

Achenbach, T.M. and Edelbrock, C.S. (1983). *Manual for the Child Behavior Checklist and the Revised Child Behavior Profile*. Burlington, VT: Thomas M. Achenbach.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31: 419–455.

Cairns, R. B. and Cairns, B. D. (1991). Social cognition and social networks: A developmental perspective. In D. J. Pepler and K. H. Rubin (eds), *The Development and Treatment of Childhood Aggression*, pp. 249–278. Hillsdale, NJ: Lawrence Erlbaum Associates.

Capaldi, D.M. and Patterson, G. R. (1987). An approach to the problem of recruitment and retention rates for longitudinal research. *Behavioral Assessment*, 9: 169–177.

Capaldi, D. M. and Patterson, G. R. (1989). *Psychometric Properties of Fourteen Latent Constructs from the Oregon Youth Study*. New York: Springer-Verlag.

Coie, J. D. and Kupersmidt, J. B. (1983). A behavioral analysis of emerging social status in boys' groups. *Child Development*, 54: 1400–1416.

Collins, L. M. and Horn, J. L. (1991). *Best Methods for the Analysis of Change*. Washington, DC: American Psychological Association.

Conger, R. D., Patterson, G. R. and Gé, X. (1995). It takes two to replicate: A mediational model for the impact of parents' stress on adolescent adjustment. *Child Development*, 66: 80–97.

Dishion, T. J., Patterson, G. R. and Kavanagh, K. A. (1992). An experimental test of the coercion model: Linking theory, measurement, and intervention. In J. McCord and R. Tremblay (eds), *The Interaction of Theory and Practice: Experimental Studies of Intervention*, pp. 253–282. New York: Guilford.

Eddy, J. M., Heyman, R. E. and Weiss, R. L. (1991). An empirical evaluation of the dyadic adjustment scale: Exploring the differences between marital "satisfaction" and "adjustment." *Behavioral Assessment*, 13: 199–220.

Farrington, D. (1986). What have we learned from major longitudinal surveys? In D. P. Farrington, L. E. Ohlin and J. Q. Wilson (eds), *Understanding and Controlling Crime: Towards a New Researching Strategy*. New York: Springer-Verlag.

Farrington, D. P. and Hawkins, J. D. (1991). Predicting participation, early onset, and later persistence in officially recorded offending. *Criminal Behaviour and Mental Health*, 1: 1–33.

Forgatch, M. S. (1991). The clinical science vortex: A developing theory of antisocial behavior. In D. Pepler and K. H. Rubin (eds), *The Development and Treatment of Childhood Aggression*, pp. 291–315. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gollob, H. F. and Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, 58: 80–92.

Nesselroade, J. R. and Baltes, P. B. (eds) (1979). *Longitudinal Research in the Study of Behavior and Development*. New York: Academic Press.

Olweus, D. (1979). Stability of aggressive reaction patterns in males: A review. *Psychological Bulletin*, 86: 852–875.

Patterson, G. R. (1988). Family process: Loops, levels, and linkages. In N. Bolger, A. Caspi, G. Downey and M. Moorehouse (eds), *Persons in Context: Developmental Processes*, pp. 114–151. Cambridge, MA: Cambridge University Press.

Patterson, G. R. (1992). Developmental changes in antisocial behavior. In R. D. Peters, R. J. McMahon and V. L. Quinsey (eds), *Aggression and Violence Throughout the Life Span*, pp. 52–82. Newbury Park, CA: Sage.

Patterson, G. R. and Bank, L. (1987). When is a nomological network a construct? In D. R. Peterson and D. B. Fishman (eds), *Assessment for Decision*, pp. 249–279. New Brunswick, NJ: Rutgers University Press.

Patterson, G. R. and Bank, L. (1989). Some amplifying mechanisms for pathologic processes in families. In M. R. Gunner and E. Thelen (eds), *Systems and Development: The Minnesota Symposia on Child Psychology*, Vol. 22, pp. 167–209. Hillsdale, NJ: Lawrence Erlbaum Associates.

Patterson, G. R. and Capaldi, D. M. (1990). A mediational model for boys' depressed mood. In J. Rolf, A. S. Masten, D. Cicchetti, K. H. Nuechterlein and S. Weintraub (eds), *Risk and Protective Factors in the Development of Psychopathology*, pp. 141–163. Cambridge: Press Syndicate of the University of Cambridge.

Patterson, G. R., Crosby, L. and Vuchinich, S. (1992). Predicting risk for early police arrest. *Journal of Quantitative Criminology*, 8: 335–355.

Patterson, G. R., DeBaryshe, B. D. and Ramsey, E. (1989). A developmental perspective on antisocial behavior. *American Psychologist*, 44: 329–335.

Patterson, G. R., Reid, J. B. and Dishion, T. J. (1992). *A Social Learning Approach: IV. Antisocial Boys*. Eugene, OR: Castalia.

Patterson, G. R. and Stoolmiller, M. (1991). Replications of a dual failure model for boys' depressed mood. *Journal of Consulting and Clinical Psychology*, 59: 491–498.

Patterson, G. R. and Yoerger, K. (1993). Developmental models for delinquent behavior. In S. Hodgins (ed.), *Crime and Mental Disorder*, pp. 140–172. Newbury Park, CA: Sage.

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade and P. B. Baltes (eds), *Longitudinal Research in the Study of Behavior and Development*, pp. 263–301. New York: Academic.

Rogosa, D. R. and Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50: 203–228.

Stoolmiller, M. (1990). Latent growth model analysis of the relation between antisocial behavior and wandering. Unpublished doctoral dissertation, University of Oregon, Eugene.

Stoolmiller, M. and Patterson, G. R. (1995). Predicting the onset and frequency of official offending for male delinquents. Unpublished manuscript available from Oregon Social Learning Center, Eugene, OR.

Stoolmiller, M. and Bank, L. (in press). Autoregressive effects in structural equation models: We see some problems. In J. Gottman and G. Sackett (eds), *The Analysis of Change*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sullivan, J. L. (1974). Multiple indicators: Some criteria of selection. In H. M. Blalock (ed.), *Measurement in the Social Sciences*, pp. 93–156. Chicago: Aldine.

This page intentionally left blank

Chapter 11

Minimizing panel attrition

Heather Laurie

1 Introduction

This chapter examines survey nonresponse and attrition in the context of longitudinal designs where individuals may be asked to take part in a survey over an extended period and be interviewed at several points in time. Current practice for minimizing attrition on longitudinal surveys is reviewed and recommendations for good practice made. While the focus is on household panel survey designs, where typically all members of the household are interviewed and followed over time, many aspects apply equally to other longitudinal survey designs. The chapter is set out in five main sections. We begin by defining attrition and briefly discuss the potential impact of attrition on data quality. Section 2 discusses the impact of survey design features on attrition in longitudinal surveys. Section 3 describes the process of attrition, the techniques which are commonly used to maximize response and minimize attrition and what is known about the success of these techniques. Section 4 examines one technique in particular by looking at existing evidence on the effect of incentives on reducing nonresponse and attrition over time. Section 5 concludes with some recommendations for good practice in the design and conduct of longitudinal surveys in order to minimize attrition.

1.1 Defining attrition

When respondents drop out of a longitudinal survey following the first round of data collection this is called attrition, i.e., losses to the sample over and above natural losses through death. Attrition is a process which leads to a cumulative reduction in the initial sample size over time. Attrition is also of concern if it is systematically related to respondent characteristics as it presents the possibility of attrition bias affecting the accuracy of our estimates in substantive analysis[1]. Sample attrition (sometimes also called panel mortality) is therefore a special case of nonresponse which applies to longitudinal surveys and is an area where survey organizations must develop and use additional types of procedures to minimize losses to the sample.

Attrition has a number of implications for data quality and the accuracy of estimates in analysis. The first issue is sample size, particularly for relatively small subgroups within the population. If cell sizes become too small the range of analysis possible with the data will be restricted. Secondly, if attrition is nonrandom from a statistically selected population,

[1] See Kaspryzk et al. (1989) and Brown et al. (1996) for a discussion of the impact of attrition bias on substantive analysis.

we can no longer be sure that the remaining sample members represent the population of interest. We risk what is known as differential or systematic attrition when those who drop out have specific demographic or other characteristics, a process which could lead to attrition bias (Kalton et al., 1990; Pannenberg and Rendtel, 1996; Neukirch, 2002; Lynn and Clarke, 2002). This is of particular concern if those characteristics are associated with the outcome of interest. For example, if people who move jobs frequently have higher levels of attrition than those in stable jobs we will underestimate the rate of job change in the population over time.

Longitudinal analysis is interested in the analysis of events, transitions and changes over time and in estimating the likelihood of those transitions or events occurring for particular types of people. While there is some evidence that much of the attrition we observe tends to be largely random or has little effect on substantive outcomes (Brown et al., 1996; Watson and Wooden, 2006), in common with nonresponse on cross-sectional surveys, there are particular groups who are more likely to drop out of a longitudinal survey than others (Kalton et al., 1989; Groves and Couper, 1998). There are many statistical techniques for adjusting for nonresponse and attrition in panel surveys and once these adjustments have been made there may be little evidence of bias across a range of estimates (Kalton et al., 1989). Arguably, panel surveys are better placed to provide accurate weights due to knowledge about respondents' characteristics and circumstances from earlier waves. However, while post-field adjustment is possible it is never ideal, with the key element in ensuring high-quality data being prevention of attrition in the first place.

Attrition is caused by three main aspects of the longitudinal survey process:

- Geographical mobility which leads to a failure to trace sample members.
- A failure to contact sample members at a known address.
- Refusal to take part in further rounds of interviewing after at least one interview.

Attrition can be defined in a number of ways and display different patterns ranging from intermittent patterns of nonresponse across a number of sweeps or "waves" of the survey to a total loss from the sample altogether[2]. Nonresponse at one wave may be an indicator of increased propensity to drop out of the survey altogether, but does not necessarily constitute attrition if the survey procedures allow the possibility that individuals could be interviewed at a later wave. The implication is that anyone who drops out of a longitudinal survey (for reasons other than death) remains potentially eligible for interview, even if the chances of gaining further interviews are low.

For the purposes of this chapter we discuss minimizing attrition in terms of total losses to the sample due to a failure to trace and nonresponse due to refusals and noncontacts, aspects which require different approaches and strategies to maximize response throughout the survey process. While survey and fieldwork procedures are central to minimizing attrition, the first element to consider is how the survey design parameters will affect likely attrition rates.

2 Longitudinal survey design and attrition

Longitudinal surveys typically follow the same individuals over an extended period of time and what all longitudinal designs have in common is that they will suffer from attrition to a greater or lesser extent. There are many designs

[2] On household panel surveys each round of interviews is typically called a "wave" while other longitudinal designs, such as cohort designs, use "sweep" for each round of data collection.

of longitudinal survey and the choice of design depends largely on the research questions being posed and the data required. If the research is interested in following children as they develop across the life course, a cohort design comprising a group of children sampled at birth with relatively infrequent data collection may be appropriate. If the interest is in income and poverty dynamics across the whole population, then a panel design with a sample drawn from the whole population and with more frequent interviews may be the best design. Buck et al. (1996) provide a good summary of the main longitudinal survey design options and the suitability of each design for collecting particular types of data.

Some survey designs allow for attrition in the original sample design, adopting some form of periodic sample refreshment or by using a rotating panel design. A rotating panel design is where sample members are systematically dropped after a certain period in the survey and new sample members recruited. The US Census Bureau's Survey of Income and Programme Participation (SIPP) and the Canadian Survey of Labor and Income Dynamics (SLID) both use a rotating panel design even though SIPP has a relatively frequent rotation pattern compared to SLID. The argument for this type of design is that it ensures the sample remains representative of the population at any given time point in the survey. Set against this are the research losses due to having a limited time span for any given sample member, making research which is interested in long-term outcomes problematic.

It is not the purpose of this chapter to review the various longitudinal survey design options but it is important to remember that the survey design will impact on attrition levels in a number of ways. As such, minimizing attrition begins with the design decisions made at the outset of the survey design process. The procedures and resources required to maintain contact with sample members and encourage cooperation over time should form an integral part of the design process. Procedures designed to minimize attrition should take into account the design parameters of the study in order to tailor those procedures most effectively to meet the needs of the survey.

2.1 Sample design and characteristics of the population

Specific characteristics of the sample population under study may be a factor in determining attrition levels (Kalton et al., 1989). The age of sample members, their lifestyles, family, housing, and employment situations will all affect not only their propensity to respond but also their likelihood of moving address (Lepkowski and Couper, 2002; Couper and Ofstedal, 2006). For example, a sample of younger people will typically be more mobile geographically than a sample of the retired population and as a result affect the ability of the survey organization to maintain contact with them. Similarly, a cohort design, where children are sampled at birth or at a particular age, may foster a greater sense of belonging and being part of a special group than a general population sample and so increase the likelihood of continuing cooperation with the survey. The British Birth Cohort Studies, for example, report high response rates after following individuals throughout their life and into their 40s and 50s (Hawkes and Plewis, 2006).

There is some evidence of between country differences in response rates, most of which are likely to be due primarily to differences in survey design and fieldwork procedures rather than intrinsic cultural differences (de Heer, 1999). Watson (2003) in an analysis of attrition in fourteen countries in the European Community Household Panel Survey (1994–2001) finds that the retention rates varied between 82% in Portugal and 57% in Ireland over the first five waves of the survey. Attrition rates are typically lower in the US than in Western European countries. On the Panel Study of Income

Dynamics (PSID) in the US just over a quarter of the sample had been lost after eight years between 1968 and 1975 (Fitzgerald, Gottschalk and Moffat, 1998) compared to the German and British household panel surveys which lost around 34% of their original samples over the first eight years of these panels and the Dutch socioeconomic panel which suffered a loss of 50% of the original sample over the same period (Watson and Wooden, 2006).

An upward trend in attrition rates in recent years has been noted (de Leeuw and de Heer, 2002). For example, the SIPP, which has a rotating panel design with quarterly interviews conducted over eight waves (32 months), lost 31% of households recruited for the 1996 panel over the 32-month survey period compared to 22% for the panel recruited in 1984 (Watson and Wooden, 2006).

For a face-to-face survey, the extent to which the sample is clustered may affect both initial response and subsequent attrition. The number of clusters in the sample is important not only for determining design effects and effective sample sizes but for ensuring sufficient coverage of interviewers across all areas being surveyed in a longitudinal survey. A clustered sample helps with the process of making the initial contact with households as interviewers can work households in the same area efficiently (Morton-Williams, 1993). As the panel progresses, the sample will de-cluster over time as households and individuals move and having trained interviewers covering all areas is important to avoid the geographically distant case being lost to the survey. For example, on the British Household Panel Survey (BHPS) individual movers to the remote Scottish Highlands and Western Isles are followed, with the costs of interviewer travel being factored into the overall survey costs (Lynn (ed.), 2006). Alternative strategies might include the use of mixed mode data collection strategies, e.g., using telephone interviews for remote cases as is done on the Household Income and Labour Dynamics in Australia survey (HILDA) (Watson and Wooden, 2004).

2.2 Frequency of interviewing and perceived burden

The frequency of interviewing set out by the survey design is likely to impact on attrition. The first aspect to consider is simply maintaining contact with respondents. With frequent interviews it is generally easier to keep in touch with sample members, whereas a design with interviews at four- or five-yearly intervals might result in higher levels of attrition as it is more difficult to maintain contact with respondents (Lepkowski and Couper, 2002; Couper and Ofstedal, 2006). However, as Kalton et al. (1989) point out, tracing respondents is time-consuming, so a longer interval between interviews can provide more time to carry out this work. Set against this is the level of burden imposed by more frequent interviews which may lead to "panel fatigue" and higher levels of attrition (Laurie et al., 1999; Kalton and Citro, 1993). Respondents take a view on the level of burden any survey implies for them, weighing up the costs and benefits of taking part from their own perspective, in other words the "opportunity costs" of taking part (Lynn et al., 2005; Groves and Couper,1998). Interviews conducted at three-monthly intervals may be seen as overly burdensome by respondents and lead to higher levels of attrition than a design with an annual or bi-annual interview, for example.

Looking at the perceived future cost of taking part in a longitudinal survey, Apodaca, Lea and Edwards (1998) found a 5% decrease in the response rate when respondents to the Medicare Current Beneficiary Survey in the US were read a statement telling them the survey was longitudinal and they would be contacted a few times each year. Even though response may be depressed as the wave respondents are told that the survey is longitudinal, there is also some evidence that overall response rates may be higher at subsequent waves (Lynn, Taylor and Brook, 1997). The decision about when and how to tell respondents they are being recruited to a longitudinal survey is therefore not only an ethical issue in terms of gaining informed consent but one which has potential implications for both the initial response rate and subsequent attrition.

2.3 Interview length and complexity

Interviews which are long or seen as having subject matter which is too personal or overly intrusive by respondents may lead to higher levels of attrition (Groves and Couper, 1998; Hill and Willis, 2001; Kalton et al., 1990). The complexity of the questionnaire and how easy respondents find it to answer may also affect attrition. As Lynn et al. (2005) note, interview length and complexity is something which respondents often say is a factor in refusing to participate at second and subsequent rounds of a longitudinal survey. While longer questionnaires are in general associated with respondents being less cooperative at subsequent contacts, the evidence on the effect of questionnaire length on attrition is somewhat mixed. Some studies have reported a decrease in attrition rates after a shortening of the questionnaire while others have found either no effect or even a positive effect of a longer interview (Lynn et al., 2005). On SIPP, respondents with a shorter interview were more likely to drop out at the subsequent round than those with a longer interview (Galvin et al., 2000). The relationship between interview length, complexity, and interest in the research topic is therefore likely to be a complex one, where a longer interview may suggest the respondent has a greater interest in the survey and is therefore less likely to drop out (Watson and Wooden, 2006).

2.4 Saliency of topic coverage

A high level of perceived saliency can be an important factor in gaining respondent cooperation not only at the first wave of a longitudinal survey but at subsequent waves (Groves et al., 2000; Dillman, 2000; Lynn et al., 2005). Saliency is the degree to which the respondent feels the survey is relevant to them, either because it is concerned with their own life experiences or has some intrinsic value for the community as a whole. In a longitudinal survey, the decision to participate in future waves is in part dependent on the experience of having taken part in a previous interview, how much the respondent enjoyed taking part and found the subject matter interesting, as well as how important the respondent sees the survey as being for the wider community or society. If the experience was enjoyable and the respondent felt that the questions were relevant to their own life, they are more likely to take part at later waves (Hill and Willis, 2001; Groves and Couper, 1998). If the research is seen as valuable for the wider community, this appeals to respondents' sense of altruism and civic duty and can increase response rates (Dillman, 2000). Designing the questionnaire to minimize complexity for respondents and maximize the perceived saliency of the research is therefore an important element in minimizing attrition and ensuring future cooperation.

2.5 Mode of data collection

The mode of data collection used will also affect attrition rates, with surveys conducted face-to-face by interviewers in respondents' homes tending to have higher response rates and lower attrition rates than those where the primary means of contact is by telephone, post, or web survey. Mixed mode data collection is often used in longitudinal surveys as a means of increasing response and offering respondents a choice of how they would prefer to respond (Lynn et al., 2005). The German Socio-Economic Panel Survey (SOEP) has used a mixed mode approach from the outset of the survey in 1984, with respondents having a choice of either a face-to-face interview or returning a self-completion questionnaire. There are also examples of surveys which have started with a face-to-face mode and shifted to either another primary mode altogether or to some combination of mixed mode data collection strategy. The PSID was initially a face-to-face survey from 1968 to 1972, moving to primarily telephone collection in 1973 and later introducing CATI (Computer Assisted Telephone Interview) technology in 1993. The Health and Retirement Survey (HRS) bi-annual interview used a mixed mode design of face-to-face and telephone interviews, with the initial baseline interview being face-to-face and follow-up interviews, primarily telephone, a strategy which changed in 2004 so that the majority of interviews were conducted face-to-face (Juster and Suzman, 1995; Couper and Ofstedal, 2006). The National Longitudinal Survey of Youth (NLSY) offers respondents the choice of telephone or face-to-face approaches and the BHPS and HILDA both use telephone interviews as part of their refusal conversion procedures when a face-to-face interview has not been possible (Burton et al., 2005; Watson and Wooden, 2004). HILDA also uses telephone interviews for respondents who move to areas outside the initial sampling points where sending an interviewer would be uneconomical (Watson and Wooden, 2004). The use of web questionnaires is widespread for consumer panels but they are not yet common on other longitudinal surveys. However, as the penetration of access to the internet increases and the problems of survey nonresponse accumulate, it is likely to be a technique that is exploited in the future[3] . Using a mixed mode design raises data quality concerns about the potential for mode effects which may bias estimates, and these need to be balanced carefully against the possible benefits of increased response (Voogt and Saris, 2005).

2.6 Following rules and sample management

Panel surveys implement a set of rules about who to follow throughout the life of the survey and how to define individuals or households as eligible at a given wave (Kasprzyk, 1989). These will typically cover the inclusion of new household members and the conditions under which sample members should be followed and remain eligible for interview. The following rules adopted for a given survey can have a direct impact on attrition in some cases.

For example, the initial sample for a survey may not include people living in institutions but may have following rules which state that anyone moving into an institution remains eligible for interview and should be followed and interviewed. Institutions vary in terms of the difficulty of gaining access to respondents with managers or others often acting as gatekeepers, e.g., the manager or matron of a residential or nursing home for the elderly. In these types of circumstances attrition may therefore be higher due to lack of access to the respondent.

Decisions about sample management and who to issue to field at each wave can have a marked effect on attrition rates. The issues are essentially to do with whether or not previous round refusals and noncontacts should be issued to field for a further attempt at subsequent waves. If all refusals and noncontacts were automatically withdrawn from the sample on the occasion of the first nonresponse, this would significantly increase attrition rates (Rodgers, 2002). Re-issuing previous noncontacts is not problematic but re-issuing refusals to field can raise ethical questions about when it is appropriate to ask people who have previously refused for another interview. Many refusals are “situational”, i.e., there is a particular circumstance at a given point in time which makes it difficult for the respondent to give an interview (Burton et al., 2005). This might be a temporary illness, bereavement, the birth of a new child, starting a new and stressful job, and

[3] Mick Couper (2000) provides a useful review of the use of web surveys.

so on. In these types of circumstances where the refusal is not an objection to taking part in the survey per se, an interview can often be achieved at a later date. Alternatively, the interviewer may simply have called on the household at a bad time and a subsequent approach may be more successful.

In the context of a longitudinal survey, assessing the likely combined effects of the survey design including the characteristics of the sample, following rules and sample management decisions, mode of data collection, frequency of interview, perceived burden, saliency, interview length, and complexity of the questionnaire need to be balanced in order to maximize response and minimize attrition at later waves of the survey while continuing to meet the data requirements of the study.

3 The process of attrition and techniques for maximizing response

The process of attrition is closely linked with theories of nonresponse and response propensity (Dillman, 2000; Groves and Couper, 1998; Lepkowski and Couper, 2002) with the key difference that respondents, having once taken part in the survey, base their future decisions on whether or not to cooperate on that experience. Dillman sees survey response as a social exchange where *". . . the actions of individuals are motivated by the return these actions are expected to bring"* (2000, p. 14). The three key elements of this exchange are rewards, cost, and trust, where the rewards are what the individual expects to gain from taking part, the cost is what one gives up in order to take part (e.g., time), and trust is the expectation that the rewards will outweigh the costs in the long run. Groves and Couper (1998) provide an alternative conceptual framework for understanding the decision to either cooperate or refuse to take part in a survey. The influences they include are the social environment, characteristics of the household(er), survey design features, interviewer attributes and behavior, and the interaction between the interviewer and householder when contact is made. Lepkowski and Couper (2002) posit three key steps in gaining cooperation for any survey. First you must locate the sample member, second you must contact the sample member, and finally you must gain the cooperation of the sample member. This approach is useful when thinking about the attrition process in longitudinal surveys as the central element of any longitudinal design is that you are following and interviewing the same people over time.

There are many standard techniques for minimizing attrition in longitudinal surveys and survey practitioners are always looking at developing new and innovative ways of maintaining their samples. How these techniques are implemented varies across surveys depending on the requirements of the survey design and the judgement of the survey practitioner about what is most appropriate for a given type of population. It is useful when designing a set of procedures to see them as a package of different elements, each of which contributes in differing ways to the overarching objective of minimizing attrition as far as possible. How these elements are tailored depends on the specific needs of the survey, with some elements being more effective in certain circumstances than others.

3.1 Keeping track of sample members and maintaining contact

At the first interview of a longitudinal survey the process of locating sample members is identical to any cross-sectional survey and will depend largely on the sampling frame and contact mode used for the survey. The difference in a longitudinal survey is the need to collect additional information at the first wave that may be needed for identifying and finding sample members at future waves of the survey. Rates of moving will vary between country and also by the type of population being surveyed, but as a general guide can be expected to range between 10% and 25% of cases a year. The propensity to move varies with individual and household characteristics such as age and stage of the lifecycle, with younger people and those in the process of family formation being more likely to move, for example. Having the ability to trace movers to a new address is therefore an essential prerequisite of locating them at the following round of interviews. For surveys using a face-to-face interview, much of the tracing will happen on the ground during fieldwork when interviewers discover someone has moved and find their new address from the current residents of the address. Where a new address cannot be found by the interviewer resources for tracing them through other means are needed.

Collecting additional contact details

A first and seemingly simple procedure is ensuring collection of full names and titles of all sample members at wave 1, including those who may not have been interviewed, such as children living in a sample household. As this is not something which is done on most cross-sectional surveys on grounds of maintaining confidentiality for respondents, it is sometimes overlooked. At the first round, and at each subsequent round of interviews, collect as much additional contact information as possible for respondents themselves, including alternative telephone numbers such as a work number and mobile telephone numbers, and an email address if they have one. Details of at least one and ideally two stable contact persons, such as a relative or close friend, should also be collected including their full name, address, telephone number(s), and the relationship of the contact person to the respondent. In a design where multiple people are being interviewed within a household it is advisable to collect different contact names for each respondent where possible. For example, if a couple separate at a later wave and one partner cannot be traced it may be difficult to contact their former in-laws or a close friend of their former partner to get a new address for them. At each wave of the survey respondents should be asked how likely it is they will move before the next interview as this can provide advance warning to the survey organization about possible moves in the future so that tracing can begin early.

Keeping in touch exercises (KITEs)

KITEs are where respondents are contacted in some way between interviews in order to encourage a feeling of belonging amongst survey respondents and to gain information about people who have moved address, have emigrated, or even died since the last interview. KITEs are usually carried out through a mailing, including a letter, a brochure of key findings, a change of address card, or other materials relevant for the survey. The KITE can also be used as a vehicle for asking respondents to confirm their current address and notify the survey organization of any moves out of the household. Returned mail for people who have moved also allows the survey organization to know that someone has moved and is likely to need tracing before the next fieldwork period begins.

While there is no experimental evidence on the effect of providing feedback to respondents in the form of some brief survey findings, anecdotally respondents report they appreciate this and it is seen as a means of fostering an understanding of the research amongst respondents (Dillman, 2000). The assumption here is that if respondents have evidence of the value of the research as well as findings that demonstrate some relevance to their own life, they will be more inclined to continue taking part. Some surveys also use birthday cards or Christmas or other religious festival cards as a means of keeping in touch and fostering loyalty to the survey. In Dillman's theoretical framework of social exchange and his "tailored design" method (Dillman, 2000), providing details to respondents or sending thank-you letters or other items showing appreciation for their participation and providing information about how the data are being used is a way of rewarding and developing trust amongst respondents.

Developing a recognizable "brand" for the survey through the use of consistent color ways or survey logos on letters, leaflets, and other survey materials are also techniques designed to encourage loyalty to the survey and help respondents feel part of the overall project.

Tracing during and between fieldwork periods

During fieldwork where a face-to-face interview is being attempted, interviewers become a major resource for tracing. Interviewers are able to enquire on the ground from current residents, neighbors, friends, and relations and are often most successful at finding new addresses for movers. Interviewers can attempt to contact respondents by telephone using all available numbers (mobiles, work numbers etc.), by checking directory enquiries and telephoning contact persons given at a previous wave by the respondent. Where tracing on the ground fails, more centralized tracing procedures take over.

Tracing respondents between fieldwork periods to ensure that address details are as up to date as possible before going into the field is referred to as a prospective "forward tracing method" by Burgess (1989). As mentioned in the previous section, this is often combined with a KITE. In order to encourage respondents to tell the survey organization when they move address between interview points, a contact point, ideally a named person, with a phone number and email address should be provided to respondents. Change of address cards which respondents can return postage-free to notify a change of address are also used by most surveys, and websites designed for respondents can also incorporate a change of address form. On some surveys, respondents are given a small incentive for returning or notifying a change of address. On the BHPS respondents are paid a £5 incentive for notifying a change of address between interviews (Laurie et al., 1999) and PSID pay $10 to respondents who return a card either verifying or updating their address between interview points. Couper and Ofstedal (2006) found that this had a significant positive effect on tracing during the 2005 PSID data collection. They found that fewer than 6% of respondents who verified their address between waves needed tracing during fieldwork, compared to 30% of those who did not return the postcard. In addition, those who returned the postcard required significantly fewer calls to contact at the following wave than those who did not return the card, representing savings in fieldwork costs as well as tracing costs.

If possible, it is also prudent to maintain a complete historical record of all previous addresses for sample members as previous co-residents of the untraced mover can be contacted and may have a new address for the sample member. Providing interviewers with the contact name details given by respondents at the previous interview so they can use these during fieldwork if they are unable to find a new address otherwise, can be a useful technique. This has the advantage of reducing time delays as interviewers are able to attempt to trace some respondents without having to contact the field office which in turn frees up staff time for tracing the more difficult cases interviewers are unable to locate.

Couper and Ofstedal (2006) argue that surveys should, where possible, collect systematic information on the steps involved in the tracing process and the outcomes of each step so that these data can be used to model the tracing process and better understand the effectiveness of the differing approaches and techniques used. Having better information on the tracing process would also allow procedures to be tailored and targeted for particular respondents or situations more efficiently than is currently the case on most surveys. If the population being surveyed is from a specific demographic or occupational group, is particularly mobile, or has not been contacted for a protracted period, then special procedures tailored to those populations should be developed (see, for example, Wright, Allen and Devine, 1995; Menedez, White and Tulsky, 2001).

The success of tracing will depend on the quality of the tracing information available, whether or not the move is long distance or not, and whether the move is a whole household or partial household move, i.e., one or more sample members have moved but others are still resident at the last known address. In the UK, Buck (2000) found that of all movers 63% were short distance (within the local authority area), 21% were medium-distance moves (between local authority districts in the same region) and the remainder were long-distance moves. In general, moves within the local area and partial household moves are easier to trace than long-distance whole household moves (Laurie et al., 1999).

Tracing using public records and linking to administrative data sources

There are a variety of publicly available administrative sources which can be used to help with tracing respondents. These include voting registers, vehicle or property registers, the web, and telephone directories, for example. Some surveys also use media appeals, something which was done on the UK Birth Cohort Studies (Lynn et al., 2005). There may also be administrative records which are not publicly available, such as welfare, benefit, immigration, defense, health, or housing records, which can be used for tracing respondents. This raises some ethical issues and it can be difficult to gain permission to access such data, but if it is possible it can be an effective means of finding sample members who cannot be traced in any other way.

Having links to death registers is also valuable as this enables the survey organization to be notified or confirm the death of a sample member. During fieldwork it can be difficult to establish whether or not a sample member is still alive and therefore whether or not they are still within the eligible population, so having links to death registers is helpful. Depending on national legal requirements, links to register databases of this type may require the permission of respondents and study protocols should be designed with this in mind if any data linkage is considered (Jenkins et al., 2004). Depending on the laws of privacy in a given country, it may also be possible to trace people through private or commercial sources such as credit records or through using private agencies such as debt collection agencies or private investigators. The web is also becoming an increasingly used resource for finding individuals, either through search engines or using commercial databases. Clearly, not all of these avenues would be free so the choice of methods will depend on the survey budget as well as any ethical and legal constraints involved.

Allowing an extended fieldwork period after the main fieldwork has been completed to enable tracing of difficult to find cases is an effective strategy. It has implications for fieldwork operations as relatively small numbers of cases remain active for an extended period but is an important strategy for minimizing losses and attrition bias, particularly as the most difficult to trace respondents tend to have particular characteristics. With suitable procedures in place tracing is usually fairly successful with location rates of over 90% for most movers. For example, on the PSID 18% of families (n = 1441) interviewed in 2003 had to be traced in 2005 with just 48 families not being successfully traced (Couper and Ofstedal, 2006). The Health and Retirement Study in the US (HRS) had to trace 11% of respondents in 2004, with just 1.3% remaining untraced. Tracing rates are slightly lower on two of the major European panel surveys. The BHPS located 94% of the 15% of sample members who needed tracing between 2003 and 2004 and the German SOEP located 96% of the 14% who needed tracing between 2003 and 2005 (Couper and Ofstedal, 2006). Most importantly, the organizational and staff resources to carry out tracing between and during fieldwork periods must be in place, something which adds to the cost of carrying out a longitudinal survey but which is also recognized as a critical element of minimizing losses to the sample.

3.2 Making contact

After the first wave of a longitudinal survey, noncontact rates are in general fairly low with mature face-to-face panel surveys reporting low noncontact rates of less than 2% in most cases. Having made contact at the first wave of the survey, the address, names of sample members, and characteristics of the household are known, all of which increase the chances of successfully contacting respondents at later waves. On longitudinal surveys the procedures for making contact include most of the standard means used for cross-sectional surveys. In particular an advance letter from the survey organization which presents the survey as a legitimate research exercise, tells respondents when to expect the interviewer to call, and provides contact numbers and assurances of confidentiality is generally seen as good practice. Specifying a minimum number of calls which interviewers must make, including the number of evening and weekend calls required before declaring a household as a noncontact, has also been shown to improve response rates (Morton-Williams, 1993; Campanelli et al., 1997). Persistence does pay off and interviewers should keep trying noncontacts throughout the fieldwork period to increase the chances of contacting people who happened to be away on holiday or business in the earlier stages of fieldwork or who are rarely at home.

One advantage of a panel or longitudinal survey is that information about when a successful call was made at the previous wave is available and can be fed forward and given to interviewers so they can plan the best time to approach the household. This is particularly important where a respondent works nights or shifts of some kind or has some specific requirements which they have notified to the survey organization, e.g., will only be interviewed on the weekend, no evening calls, must be telephoned in advance for an appointment, etc. Interviewers should be encouraged to maintain as much flexibility as possible during fieldwork in order to fit in with the schedule of the respondent, even if this means making multiple calls to the household to interview different household members. Keeping records of any contacts with respondents between interview points so that these can be passed to interviewers where necessary can also help with the process of making contact. For example, if news of a bereavement, serious illness, or holiday dates has been passed to the survey organization these are useful pieces of information for the interviewer to know before they attempt the household and can increase their chances of making contact most efficiently. In the setting of a centralized telephone unit, the call pattern is determined by the system which again should use any previous information to tailor the approach as far as possible (Bennett and Steel, 2000).

3.3 Gaining and maintaining cooperation

Using experienced, well-trained interviewers is clearly important for maximizing response rates on any survey. For longitudinal surveys, interviewers need to be trained in additional procedures such as tracing movers as well as response maximization techniques designed to maintain complete response histories for as many respondents as possible. The approach to the household and the ability to explain the purposes of the survey, to respond effectively to any concerns expressed by respondents and to convey an understanding of the importance of the longitudinal design are central elements in gaining respondent trust and commitment to the survey (Campanelli et al., 1997). As Groves and Couper (1998) argue, the interviewer must be able to tailor their approach to the individual respondent, something which is facilitated by having prior knowledge of the circumstances of individuals and households. Interviewers must also be sensitive enough to recognize when they should retreat from a household before a refusal hardens (Morten-Williams, 1993). Using previous response histories of sample members to model who are most likely to refuse at the following wave so that targeted strategies can be developed for these cases can also help to minimize losses.

For many respondents, the interview is seen as an enjoyable event and it is not uncommon on longitudinal surveys for respondents to be waiting for the interviewer to call. There is some evidence that maintaining continuity of interviewers has a positive effect on maintaining response rates. Hill and Willis (2001) found that having the same interviewer was the most significant factor predicting future response with roughly a 6% increase in response rates where this was the case. The argument is that interviewers and respondents build up a rapport which encourages participation and that when the interviewer changes, this link is broken to some extent. So there may be a loyalty and trust that is built on the personal relationship between the interviewer and respondent which is independent of the level of trust the respondent has for the organization carrying out the survey. In contrast, Campanelli and O'Muircheartaigh (2002) found that once area effects are properly controlled there is no interviewer continuity effect on response, even though significant variation in the continuity effect remained among refusing individuals, possibly due to unmeasured aspects of the interviewer such as their skill on the doorstep. Nonetheless, interviewer continuity is a strategy which is employed by most of the major household panel surveys which use face-to-face interviewing. Where telephone interviews are used, this type of continuity is not normally possible as interviewer turnover rates tend to be higher within telephone units. Even if it were possible, it may not have any effect on response as the telephone contact is by definition less personal and more anonymous than a face-to-face contact (Budowski and Scherpenzeel, 2005).

As discussed in the previous section, ensuring interviewers maintain flexibility in terms of making multiple calls at the same household to complete all interviews and being prepared to fit in with the circumstances of individual respondents is important. Close monitoring of fieldwork progress has been shown to increase response rates (de Leeuw and de Heer, 2002). This may be due to the fact that interviewers are aware their performance is being monitored, but monitoring also allows early warning of any problems during fieldwork so that these can be addressed quickly and not left until the end of fieldwork when it may be too late to intervene in any way. Relevant information held about the household or respondent by the survey organization should be provided to interviewers to help with the approach to the household and also be provided during fieldwork if the respondent notifies a change or difficulty of some kind, e.g., a house move or illness. Interviewers should also be asked to provide their own assessment of how the interview went, how cooperative the respondent seemed to them, and in the case of refusals, the reasons given by the respondent for refusing, information which is important for decisions about refusal conversion and whether the sample member should be withdrawn from the sample or attempted at a later wave.

The survey should also be designed to collect what Couper (1998) terms paradata, which can include information about the characteristics of the interviewer, observations of the local area, observations about physical barriers preventing access to the address, and details of not only call patterns but of the interaction with respondents at each call. These data can then be used to develop what Groves and Heeringa (2006) have recently termed "responsive survey designs" where paradata are actively used during data collection to assess when each phase of the survey process has reached its capacity in terms of response and what strategies or additional features could be used to increase response for the remaining sample. This might include having the most productive interviewers attempt the most difficult cases, a specialized refusal conversion program or increased incentives for respondents, for example.

Refusal conversion

Refusal conversion is a common practice on both cross-sectional and longitudinal surveys but arguably is more important in the context of a longitudinal survey where the cumulative response rate over time is a critical quality indicator for the survey. Refusal conversion is where respondents who have initially refused to take part in the survey are contacted again during the current interviewing period to see if there is anything that can be done to encourage participation. Respondents refuse for a variety of reasons, some of which may be able to be catered for in some way. For example, if a respondent says that they cannot take part due to being too busy it may be possible to have the interviewer call at a time that suits the respondent. If the respondent did not like the interviewer who called on them for some reason or would prefer an interviewer with a different gender, for example, this can usually be catered for by the survey organization. Refusal conversion can be carried out face-to-face or by other means such as telephoning the respondent. Refusals will be due to a range of reasons, from simply not being interested in the survey to a specific situation such as illness or bereavement (Burton et al., 2005). Specialist interviewers who are trained in refusal conversion should deal with these cases and having sufficient information about the circumstances of the refusal will help with tailoring the subsequent approach. Of course a proportion of respondents will refuse completely to have any further contact and such wishes should be respected and the respondent withdrawn from the sample at that point.

As noted in Section 2.5 using mixed mode data collection can be useful for response maximization and if used during refusal conversion can offer respondents an alternative means of completing the interview which may suit them better. Offering respondents a choice of mode can enable the respondent to be kept within the interviewed sample and there is evidence that such strategies are effective in maintaining overall sample numbers over extended timeframes (Burton et al., 2005).

4 The use of incentives to minimize attrition

There is an extensive literature on the use of incentives in cross-sectional surveys but less in the context of longitudinal surveys[4]. The cross-sectional evidence shows that cash incentives are effective in increasing response, even though this varies by survey mode and the type of incentive strategy used. Pre-paid monetary incentives given unconditionally in advance of the interview are the most effective in increasing response compared to a monetary incentive which is dependent on response or a gift. And any incentive is better than no incentive (Church, 1993; Singer et al., 1999; James and Bolstein, 1992; Couper et al., 2005).

Incentives are more effective at increasing response on surveys which are burdensome (Lynn and Sturgis, 1997) and are also known to be more effective on surveys which typically have lower response rates and where the saliency of the research may not be high for respondents (Groves et al., 2000). Incentives work primarily by reducing refusals and have little effect on noncontact rates (Singer et al., 1999). One concern when using incentives is whether data collected from respondents who

[4] Laurie and Lynn (2006) provide a useful summary of what is known about incentives in both cross-sectional and longitudinal contexts and current practice on some of the major panel surveys.

may be less cooperative is of poor quality. However, there seems to be little effect on data quality in terms of sample composition and response distributions where either cash or a gift is offered (Couper et al., 2006). Nor do monetary incentives appear to adversely affect data quality as measured by the levels of item nonresponse or the effort expended in the interview measured by the number of words given to verbatim items (Singer, Van Hoewyk and Maher, 1998; Willimack et al., 1995).

On longitudinal surveys, in the absence of experimental evidence, it is difficult to disentangle the effect of the incentive from the survey procedures discussed in previous sections, some of which may have significant impacts on response rates. Some experimental evidence exists but the results are somewhat mixed. Overall, current evidence suggests that incentives can be effective in reducing attrition over multiple waves of a survey, and that making changes through introducing an incentive, offering higher amounts and targeting of various kinds does increase response even though these effects vary depending on the survey context. As with cross-sectional surveys, pre-paid, unconditional monetary incentives are most effective in increasing response, an effect which holds across multiple waves (James, 1997; Mack et al., 1998). However, the incentive needs to be sufficiently high to reduce attrition over time, with some evidence that smaller monetary incentives have no effect over the longer term (Mack et al., 1998). Others have found there is a positive and enduring effect on subsequent wave response for nonmonetary incentives, where entry into a lottery was offered during the life of a survey (Scherpenzeel et al., 2002). There is also some evidence of lower levels of item nonresponse where incentives are used on longitudinal surveys and a reduction in interviewer effort in terms of the number of calls required (James, 1997; Mack et al., 1998). Incentives appear to have a differential effect by demographic characteristics, with those on low incomes, with low educational qualifications and from ethnic minority backgrounds responding to the incentive more than other groups (Mack et al., 1998). Increasing the amount of incentive paid during the life of a panel has also been shown to increase response rates, in part by giving a tangible signal to respondents that their continued participation is appreciated (Laurie and Lynn, 2006).

Targeting strategies have been found to be effective, especially where previous refusals have been offered an incentive to take part at a later wave (Martin et al., 2001; Kay et al., 2001; Rodgers, 2002). One-off, large payments or “end game” payment strategies to increase response from the least cooperative sample members at the first wave of a longitudinal survey have also been used with some success (Juster and Suzman, 1995). Even though we do not know how successful these are in delivering long-term commitment to the survey, one study suggests that a large payment at the first wave had no effect on increasing or decreasing later response relative to others who initially refused and were persuaded to take part by other means, nor did the large incentive at wave 1 induce an expectation that large incentives would be offered in later waves of the panel (Lengacher et al., 1995). Targeting raises issues of equity and fairness to respondents, who may react negatively if they know that other sample members are receiving more than themselves, even though the evidence from one study suggests that this is not necessarily problematic (Singer, Groves and Corning, 1999). However, in the context of a household survey where all members are being interviewed, the use of differential incentive payments may be problematic. This is an area that deserves further enquiry as there may be unintended consequences of perceptions of inequity and maintaining the goodwill of survey respondents is paramount.

There are many areas where we have limited knowledge about the longer term effects of incentives even though some studies suggest there may be long-term beneficial effects on reducing attrition and potential bias through using incentives. However, further experimental work to establish the effect of incentives on longer term attrition, sample composition and data quality, the best targeting strategies to use, and the effect of introducing, increasing or changing the way incentives are delivered during the life of a longitudinal survey is still required. The use of incentives inevitably incurs a direct survey cost, so the likely benefits in reducing attrition rates need to be weighed carefully against the costs.

5 Conclusion and best practice guidelines

Attrition is a concern for any longitudinal survey, with high levels of attrition having the potential to significantly affect data quality and the long-term viability of the survey. Attrition rates are commonly used by data analysts and funders as a measure of the success of a survey and a critical indicator of survey quality. Any longitudinal survey design needs to include a package of procedures to minimize attrition together with the necessary resources required to implement these. While surveys vary in the techniques they use to minimize attrition there are a number of elements which should be considered, including:

- the impact the survey design is likely to have on attrition such as the frequency of interviewing, the length and complexity of the questionnaire, the subject matter of the survey, and the mode of data collection
- questionnaire design to minimize complexity for respondents, maximize saliency, maintain the interest of respondents, and make the interview an enjoyable experience
- tailoring survey procedures to suit the characteristics of the sample and population being surveyed and designing a package of measures to minimize attrition which are most appropriate to the needs of the survey
- the following of rules for the survey and the decisions about sample management, particularly the re-issuing of previous refusals and noncontacts where possible as these can have a significant impact on attrition
- how to introduce the purpose of the survey to respondents, in particular the longitudinal nature of the survey and what will be expected of respondents who agree to take part
- collection of sufficient contact details at the first interview to facilitate later tracing, including full names of all sample members, contact names for family or friends outside the household, telephone numbers including mobile numbers and email addresses
- procedures to trace respondents between and during fieldwork and for maintaining complete historical records of all previous addresses for sample members
- investigating and exploiting the full range of tracing avenues available, including publicly available records such as voter registers and directory enquiries as well as other administrative records and commercial databases where possible
- updating address records between interview points to increase the chances of making contact, reduce the amount of tracing that needs to be done during fieldwork, and reduce the number of calls required by interviewers
- the use of Keeping in Touch Exercises (KITEs) between interviewing points to maintain contact with respondents and a vehicle for them to provide information about moves or other changes in their circumstances (e.g., a recent bereavement or illness)
- thanking respondents for their participation and providing them with details of findings from the survey to develop trust, loyalty, and commitment to the survey
- "branding" of survey materials to encourage a sense of belonging to the survey
- providing respondents with a named contact person, telephone number, email address and

change of address cards so they can contact the survey organization with queries or notify changes in their circumstances

- recording details of all contacts with respondents between interview points so that relevant information can be passed to interviewers for the next round of fieldwork
- a dedicated website for respondents with information about the survey and including a change of address form and a feedback comments field
- collection of call record data which can be fed back to interviewers at the following round to improve the chances of making contact and reduce the number of calls required
- collecting systematic data on the tracing process itself to allow more efficient tailoring of procedures
- ensuring interviewers are experienced and well trained in tracing techniques, strategies for contacting respondents, and maximizing response
- ensuring interviewers are flexible throughout the fieldwork period, prepared to fit in with the needs of respondents, and keep trying noncontacts until the end of the fieldwork period
- in a face-to-face survey, ensuring interviewer continuity where possible
- setting up systems to monitor fieldwork closely so that any problems can be detected early
- allowing an extended fieldwork period to trace movers and carry out refusal conversion
- the use of mixed mode data collection strategies to maximize response
- a refusal conversion program during fieldwork which collects systematic data about the refusal conversion process so that these data can be used to model the most successful strategies for particular types of respondents
- the use of monetary and/or nonmonetary incentives and whether it is practical to target these in any way.

References

Apodaca, R., Lea, S. and Edwards, B. (1998). The effect of longitudinal burden on survey participation. 1998 Proceedings of Survey Research Methods Section of the American Statistical Association, pp. 575–580.

Bennett, D. J. and Steel, D. (2000). An evaluation of a large-scale CATI household survey using random digit dialling. *Australian and New Zealand Journal of Statistics*, 42(3): 255–270.

Brown, C., Duncan, G. J. and Stafford, F. P (1996). Data Watch: The Panel Study of Income Dynamics. *Journal of Economic Perspectives* 10(2): 155–168.

Buck, N. H., Ermisch, J. F. and Jenkins, S. P. (1996). Choosing a longitudinal survey design: The issues. Occasional Paper 96-1, Institute for Social and Economic Research, University of Essex.

Budowski, M. and Scherpenzeel, A. (2005). Encouraging and maintaining participation in household surveys: The case of the Swiss Household Panel, ZUMA-Nachrichten, 56, Jg 29 (May): 10–36.

Burgess, R. D. (1989). Major issues and implications of tracing survey respondents. In D. Kasprzyk et al. (eds), *Panel Surveys*. New York: Wiley.

Burton, J., Laurie, H. and Lynn, P. (2006). The long-term effectiveness of refusal conversion procedures on longitudinal surveys. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 169(3): 459–478.

Campanelli, P., Sturgis, P. and Purdon, S. (1997). *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates*. London: National Centre for Social Research.

Campanelli, P. and O'Muircheartaigh, C. (2002). The importance of experimental control in testing the impact of interviewer continuity on panel survey nonresponse. *Quality and Quantity*, 36(2): 129–144.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57(1) (Spring): 62–79.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. Proceedings of Survey Research Methods Section of the American Statistical Association, pp. 41–49.

Couper, M. P. (2002). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64(4) (Winter): 464–494.

Couper, M. P. and Ofstedal, M. B. (2006). Keeping in contact with mobile sample members. Paper presented at the Methodology of Longitudinal Surveys Conference, University of Essex.

http://iser.essex.ac.uk/ulsc/mols2006/programme/details.php

Couper, M. P., Ryu, E. and Marans, R. W (2006). Survey incentives: Cash vs in-kind; face-to-face vs mail; response rate vs nonresponse error. *International Journal of Public Opinion Research*, 18 (1): 89–106.

De Heer, W. (1999). International response trends: Results of an international survey. *Journal of Official Statistics*, 15(2)**:** 129–142.

De Leeuw, E. and de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little (eds) (2002), *Survey Nonresponse*, New York, Wiley.

Dillman, D. A. (2000). Mail and internet surveys. The Tailored Design Method, 2nd edn. New York: Wiley.

Fitzgerald, J., Gottschalk, P. and Moffit, R. (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *The Journal of Human Resources*, 33(2): 251–299.

Galvin, L. N., Sae-Ung, S. and King, K. (2000). Effect of interview length and proxy interviews on attrition to the Survey of Income Program Participation. Proceedings of Survey Research Methods Section of the American Statistical Association, pp. 636–640.

Groves, R. M. and Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., Dillman, D. A., Eltinge, J. L. and Little, R. J. A. (eds) (2002). *Survey Nonresponse*. New York: Wiley.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: Tolls for actively controlling survey errors and costs. *Journal of the Royal Statistical Society*, *Series A*, 169(3): 439–457.

Groves, R. M., Singer, E and Corning, A. (2000). Leverage-saliency theory of survey participation: Description and illustration. *Public Opinion Quarterly*, 64: 299–308.

Hawkes, D. and Plewis, I. (2006). Modelling nonresponse in the National Child Development Study. *Journal of the Royal Statistical Society, Series A*, 169(3): 479–491.

Hill, D. H. and Willis, R. J. (2001). Reducing panel attrition: A search for effective policy instruments. *Journal of Human Resources*, 36(3) (Summer): 416–438.

James, J. M. and Bolstein, R. (1992). Large monetary incentives and their effect on mail survey response rates. *Public Opinion Quarterly*, 56(4) (Winter): 346–361.

James, T. L. (1997). Results of the wave 1 incentive experiment in the 1996 Survey of Income and Program Participation. 1997 Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 834–839.

Jenkins, S. P., Cappellari, L., Lynn, P., Jäckle, A. and Sala, E. (2004). Patterns of consent: evidence from a general household survey. Working Paper 2004-27, Institute for Social and Economic Research, University of Essex.

Juster, F. T. and Suzman, R. M. (1995). An overview of the Health and Retirement Study, *Journal of Human Resources*, 30 (Suppl), S7–S56.

Kalton, G. and Citro, C. F. (1993). Panel surveys: Adding the fourth dimension. *Survey Methodology*, 19(2): 205–215.

Kalton, G., Kasprzyk, D. and McMillen, D. B. (1989). Nonsampling errors in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Surveys,* pp. 249–270. New York: Wiley.

Kalton, G., Lepkowski, J. M., Montanari, G. E. and Maligalig, D. (1990). Characteristics of second wave nonrespondents in a panel survey. 1990 Proceedings of Survey Methods Research Section of the American Statistical Association, pp. 462–467.

Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. P. (eds) (1989). *Panel Surveys*. New York: Wiley.

Kay, W. R., Boggess, S., Selvavel, K. and McMahon, M. F. (2001). The use of targeted incentives to reluctant respondents on response rate and data quality. Proceedings of the Annual Meeting of the American Statistical Association, Aug 5–9.

Laurie, H. and Lynn, P. (2006). The use of incentives on longitudinal surveys. Paper presented at the Methodology of Longitudinal Surveys Conference, University of Essex, July. http://iser.essex.ac.uk/ulsc/mols2006/programme/details.php

Laurie, H., Smith, R. and Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15(2): 269–282.

Lengacher, J. E., Sullivan, C. M., Couper, M. P. and Groves, R. M. (1995). Once reluctant, always reluctant? Effects of differential incentives on later survey participation in a longitudinal study. Survey Research Centre, University of Michigan.

Lepkowski, J. M. and Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little (eds), *Survey Nonresponse*, pp. 259–272. New York: Wiley.

Lynn, P. (ed.) (2006). Quality profile: British Household Panel Survey waves 1 to 13: 1991–2003. Institute for Social and Economic Research, University of Essex.

Lynn, P., Buck, N., Burton, J., Jäckle, A. and Laurie, H. (2005). A review of methodological research pertinent to longitudinal survey design and data collection. Working Paper 2005-29, Institute for Social and Economic Research, University of Essex.

Lynn, P. and Clarke, P. (2002). Separating refusal bias and noncontact bias: Evidence from UK national surveys. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 51(3): 319–333.

Lynn, P. and Sturgis, P. (1997). Boosting survey response through a monetary incentive and fieldwork procedures: An experiment. *Survey Methods Newsletter*, National Centre for Social Research, London, 17(3): 18–22.

Lynn, P., Taylor, B. and Brook, L. (1997). Incentives, information and number of contacts: Testing the effects of these factors on response to a panel survey. *Survey Methods Newsletter*, 17(3): 7–12.

Mack, S., Huggins, V., Keathley, D. and Sundukchi, M. (1998). Do monetary incentives improve response rates in the survey of income and program participation? Proceedings of the American Statistical Association, Survey Research Methods Section, pp. 529–534.

Martin E., Abreu D. and Winters, F. (2001). Money and motive: Effects of incentives on panel attrition in the survey of income and program participation. *Journal of Official Statistics*, 17(2): 267–284.

Menendez, E., White, M. C. and Tulsky, J. P. (2001). Locating study subjects: Predictors and successful search strategies with inmates released from a US county jail. *Controlled Clinical Trials*, 22(3): 238–247.

Morton-Williams, J. (1993). *Interviewer Approaches*. Aldershot: Dartmouth Publishing.

Neukirch, T. (2002). Nonignorable attrition and selectivity biases in the Finnish subsample of the ECHP: An empirical study using additional register information. Chintex Working Paper No. 5. http://www.destatis.de/chintex/download/paper5.pdf

Pannenberg, M. and Rendtel, U. (1996). Documentation of sample sizes and panel attrition on the German Socio-Economic Panel. Discussion Paper No.137, DIW, Berlin.

Rodgers, W. (2002). Size of incentive effects in a longitudinal study, presented at the 2002 American Association for Public Research Conference, Mimeo, Survey Research Centre, University of Michigan, Ann Arbor.

Scherpenzeel, A. et al. (2002). Experimental pretest of the biographical questionnaire. Working Paper 5-02 of the Swiss Household Panel Survey, Neuchatel.

Singer, E., Groves, R. M. and Corning, A. D. (1999). Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *Public Opinion Quarterly*, 63(2) (Summer): 251–260.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T. and McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15 (2): 217–230.

Singer, E., Van Hoewyk, J. and Maher, P. (1998). Does the payment of incentives create expectation effects? *Public Opinion Quarterly*, 62(2): 152–164.

Taylor, S. and Lynn, P. (1996). England and Wales Youth Cohort Study (YCS): The effect of time between contacts, questionnaire length, personalisation and other factors on response to the YCS. Department for Education and Employment Research Paper No. 8, Sheffield.

Voogt, R. J. J. and Saris, W. E. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics*, 21(3): 367–387.

Watson, D. (2003). Sample attrition between waves 1 and 5 in the European Community Household Panel. *European Sociological Review*, 19(4): 361–378.

Watson, N. and Wooden, M. (2004). Wave 2 survey methodology. Technical Paper of the Melbourne Institute of Applied Economic and Social Research No. 1/04.

Watson, N. and Wooden, M. (2006). Identifying factors affecting longitudinal survey response. Paper presented at the Methodology of Longitudinal Surveys Conference, University of Essex, July. http://iser.essex.ac.uk/ulsc/mols2006/programme/details.php

Willimack, D. K., Schuman, H., Pennell, B-E. and Lepkowski, J. M. (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *Public Opinion Quarterly*, 59(1) (Spring): 78–92.

Wright, J. D., Allen, T. L. and Devine, J. A. (1995). Tracking non-traditional populations in longitudinal studies. *Evaluation and Program Planning*, 18(3): 267–277.

Chapter 12

Nonignorable nonresponse in longitudinal studies

E. Michael Foster and Anna Krivelyova

1 Introduction

Attrition represents one of the most serious threats to both the external and internal validity of research in psychology and psychiatry. For example, the internal validity of a clinical trial may be compromised if those lost to follow-up differ systematically from those remaining in the study. This problem is especially acute if the nature of this process differs between the treatment and control groups. The external validity of the study may be threatened as well. Those individuals remaining in the study may be rather unrepresentative of the original sample. Even if the parameter estimates have the interpretation we want for those participating (e.g., an unbiased estimate of the effect of the treatment), that estimate may not describe the experiences of the larger population. For example, a longitudinal evaluation may have trouble retaining low-income participants. In that case, the treatment effect estimated using those with complete data may not apply to the dropouts. (This problem would occur if economic status moderated the impact of the intervention or treatment.)

Fortunately, missing data is an active area of research in statistics and social science methodology (Little and Rubin, 2002; Schafer, 1997), and these methods are gradually influencing practice. Applied researchers increasingly recognize that practices common in the past often have a rather dubious statistical foundation. For example, the practice of simply replacing missing data with the mean distorts the relationship between that variable and other variables in the model or analysis. While better, conditional mean imputation (i.e., imputing the mean for similar individuals) has problems as well. That method, however, fails to account for the variation in the predicted value *within* the subgroups defined by the variables used to group observations.

On the other the hand, some procedures are appropriate in some circumstances but not others. For example, suppose an analyst is interested in three variables, Y, X and Z. Suppose that the likelihood of missing data depends on all three variables (even controlling for the other two). In this case, regression of Y on X and Z using listwise deleted data would produce estimates with desirable statistical properties as long as the missing data mechanism is “missing at random” (MAR). Described in more detail below, MAR means that among individuals with the same values of X and Z, those with missing and complete data do not differ in their values of Y. In other words, the likelihood of missing data does not depend on Y conditional on X and Z.

Other analyses of these same data require stronger assumptions. For example, suppose one just wanted to estimate the correlation between Y and X. In that case, a researcher using listwise deletion would have to assume the missing data mechanism is “missing completely at random” (MCAR)—in effect, the available data would have to represent a simple random sample of the complete and incomplete data. If the likelihood of missing data depended on Z and Z was related to X or Y, then the correlation would not accurately describe the relationship between those two variables.

Much of the methodological work on missing data assumes MAR. Multiple imputation, for example, begins with the assumption that the data are missing at random (Schafer, 1997; 1999; Schafer and Graham, 2002). Such methods are very useful, but in some cases one suspects that the data are not MAR. In a regression context, MAR would not hold if the outcome variable differed between individuals who do and do not provide data *conditional* on the explanatory variables in the model. In that case, the missing data mechanism is said to be “missing not at random” (MNAR).

Efforts to assess and correct data for MNAR are inherently somewhat speculative. After all, a full assessment of MNAR would depend on comparisons of the outcome data for those who do and do not provide data. Of course, by definition, that information is not available for the latter. Determining whether the data are MNAR, therefore, depends largely on the judgement of the researcher. That judgement depends on his or her knowledge of the outcome of interest and the process that shapes it. For example, in a longitudinal evaluation of a delinquency prevention program, the researcher may know that individuals who are incarcerated were generally unwilling to participate in the research study.[1] In that instance, that the respondents and nonrespondents would differ in terms of key outcomes, such as delinquency, seems quite likely. As we illustrate below, the data can provide some clues about whether the data are MNAR, but ultimately the researcher’s judgement is the key factor.

Statistical work in this area falls into two strands—selection models and pattern mixture models. The former is the subject of a large literature in econometrics, and we only briefly review it here. Our focus here is to describe and illustrate the latter, pattern mixture models. Our analyses rely on longitudinal data from a large evaluation of service delivery in children’s mental health services.

2 MAR, pattern mixture and selection models

This section begins by reviewing the technical definition of missing at random.

2.1 A brief review of missing at random

Following the notation in key texts, such as Little and Rubin or Schafer (Little and Rubin, 2002; Schafer, 1997), the analyst collects data and observes Y_{obs} and a pattern of responses and missing data, R. The latter allows us to partition the complete data, Y, into the former and missing data, $\mathrm{Y}_{\mathrm{miss}}$. The likelihood function for θ and ξ can be written as

$$L_{full}(\theta, \xi | Y_{obs}, R) \alpha\ f(R, Y_{obs} | \theta, \xi) \qquad (1)$$

where f is the joint probability distribution for the observed data (Y_{obs}) and response indicator (R). θ characterizes the distribution of Y_{obs}; ξ characterizes the distribution of R and

[1] And when they do, their ability to commit additional crimes is diminished while they are locked up. In that instance, the data of real interest—what crimes would they commit while in the community—is effectively missing.

generally represents nuisance parameters. In most instances, we are interested in θ and want an estimate that describes the behavior of all individuals, including those who did not provide data.

One can rewrite equation (1) as

$$f(R, Y_{obs}|\theta, \xi) = \int f(R, Y|\theta, \xi) dY_{miss} \tag{2}$$

Working with the right-hand side of the equation requires assumptions about the distribution of the missing data (e.g., that the missing data have the same distribution as Y_{obs}). Given that Y_{miss} are unobserved, checking such an assumption is difficult or impossible, and the resulting parameter estimates are often rather sensitive to that assumption.

Estimation is simplified if the data are "missing at random". In that case,

$$f(R|Y_{obs}, Y_{miss}, \xi) = f(R|Y_{obs}, \xi) \tag{3}$$

for all values of ξ and Y_{miss} evaluated at the observed values of R and Y_{obs}. In that case, equation (1.2) can be simplified to the following:

$$f(R, Y_{obs}|\theta, \xi) = f(R|Y_{obs}, \xi) f(Y_{obs}|\theta) \tag{4}$$

As a result, the likelihood function can be partitioned into two pieces—the first involving the parameter of interest θ; the second involves the nuisance parameter, ξ. Combined with an additional assumption (that θ and ξ are distinct), inferences about θ can ignore the missing data mechanism (the probability of response). In this case, the missing data mechanism is said to be "ignorable". To be clear, in this instance, one can analyze the available data as if they were complete, at least for some purposes.

A fair bit of confusion surrounds the meaning of "missing at random" in practice. MAR can accommodate a wide variety of missing data mechanisms. For example, individuals who do and do not provide data may differ quite dramatically—evidence that such differences exist is not evidence that MAR does not apply. Rather, the key issue for MAR is that the likelihood of response does not depend on Y_{miss}. MAR requires only that the available data represent a random sample of all values within subclasses defined by Y_{obs} (Schafer, 1997). In a multivariate regression, MAR means that the likelihood of missing data can depend on the covariates; in a longitudinal context, missingness may even depend on past values of the dependent variable. In intervention studies and clinical trials, the likelihood of response can vary with treatment status. Treatment-control differences in the response rate is not inherently problematic for analyses of treatment impact (Foster and Bickman, 1996).

A key wrinkle complication surrounding MAR is that one's ability to ignore the missing data mechanism may depend on the type and purposes of the analysis. For example, suppose that in a study of black–white differences in earnings, more educated workers are less likely to participate. In that case, descriptive statistics of mean earnings will be incorrect for both black and white workers. However, regression analyses of the between-group differences may be correct, as long as education is included as a regressor.[2] Within subgroups defined by race and education, MAR applies if those who respond do not differ systematically from those who do not.

Furthermore, the analyst has some control over the applicability of MAR. A regression, for example, with more covariates will extend the reach of MAR; as more covariates are added, patterns of missingness related to the added covariates are brought under the reach of MAR. Intuitively, as one conditions on more and more variables, the likelihood that the missing data mechanism represents a form of random sampling *within classes defined by the covariates*

[2] And of course, the model for earnings is specified correctly.

increases. Similarly, a fixed-effects analysis is robust to a broader array of missing data mechanisms than is a random-effects estimates model (Verbeek and Nijman, 1992).

As noted, however, missing at random (MAR) still may not apply, even when an extensive range of covariates have been added. In those instances, the unobserved values of the outcome of interest may directly affect the likelihood of response. Alternatively, the covariates may not capture completely shared predictors of the outcome and response. For example, in a longitudinal study of delinquency, individuals with the greatest propensity to offend may be less likely to provide data—they may be unwilling to provide data on criminal activities or they may be incarcerated (Foster, Fang and Conduct Problems Prevention Research Group, 2004). In that case, the data are said to be "missing not at random" (MNAR).

2.2 Alternative MNAR approaches: Selection models

Under MNAR, the likelihood function cannot be partitioned as

$$f(R, Y|\theta, \xi) = f(Y|\theta)\ f(R|Y, \xi) \tag{5}$$

The likelihood depends on a model of the complete data (i.e., *both* Y_{miss} and Y_{obs}) and a model predicting the likelihood of response. The latter depends explicitly on both the observed and missing values of the dependent variable. In terms of implementation, obtaining estimates θ and ξ involves estimation of equations predicting missingness and the outcome of interest. However, without further assumptions, the model is not identified: the observed data could be explained by any of a multitude of models of Y and of R. One identifying assumption might involve the distribution of the Y variables (e.g., normality). Another means involves an exclusion restriction—a variable that affects response but not the outcome directly. For example, in a longitudinal study, one might use characteristics of prior interviews, such as their length (e.g., Lillard and Panis, 1998).

Estimation allows for the interdependency between nonresponse and that outcome, either by allowing participation to predict the level of the outcome directly or, as is common in the econometric literature, allowing a correlation between the unobserved determinants of nonresponse and the outcome. Prior to the wide accessibility of maximum likelihood estimation, economists estimated selection models in two steps. First, one would estimate a probit equation predicting nonresponse. The results of that equation were used to calculate the inverse mills ratio, a nonlinear function of the probability of participation. In the second stage, the analyst includes that ratio as an explanatory variable in a suitable model used to predict the outcome (Wooldridge, 2002, p. 550).

In essence, selection models replace the MAR assumption with an alternative assumption. In many instances, that assumption may be more plausible than MAR. Regrettably, the results of these analyses are potentially sensitive to these assumptions, which are untestable. The best one can do is to examine the sensitivity of key findings to alternative parameterizations. Diggle and Kenward (1994) present a selection model in a well-known example involving milk production and mastitis among dairy cattle. The authors allow the likelihood of missing data to depend on the dependent variable, milk production. The model is identified by the assumed normality of the milk production variable. In the original analysis, the authors rejected MAR—they found that unobserved values of the dependent variable affect the likelihood of response (Diggle and Kenward, 1994). Kenward (1998), however, determined that the findings were quite sensitive to distributional assumptions and to the handling of two outlier cases. In particular, if the two cases were removed or an alternative distribution was used for the outcome of interest, a MAR model fit the data as well as the MNAR model fit (Kenward, 1998). This sensitivity is common and should be carefully examined (Little and Rubin, 2002). For an example, see Foster, Fang and Conduct Problems Prevention Research Group (2004). Arguing that the identification assumptions are often obscure, statisticians often eschew selection models and have developed pattern mixture models as an alternative.

2.3 Alternative MNAR approaches: pattern mixture models

The pattern mixture factors the likelihood function differently than does the selection model. In particular, this model allows the response variable to depend explicitly on the pattern of response (rather than vice versa):

$$f(R, Y|\delta, \phi) = f(Y|R, \delta)\ f(R|\phi) \tag{6}$$

Obtaining estimates of the ϕ and δ involves dividing observations into groups defined by study participation. For example, one might group observations according to the number of waves of data collection in which an individual participated. Then one estimates the outcome model of interest separately for each of those subgroups (i.e., conditioning on the response patterns).

Advocates of this approach favor it over selection models not because the model does not depend on assumptions but because those assumptions are more transparent. In particular, one estimates parameters for alternative groups of observations defined by their missing data patterns. For some combinations of some parameters and groups of data, estimating the model is not possible. For example, suppose that one is analyzing the effect of study dropout on a dataset that offers as many as six waves of data (like the empirical example below). The substantive model of interest relies on a standard growth curve to explain trends over time in the outcome of interest. Suppose that the model includes a quadratic time term, allowing for a nonlinearity in growth over time. Following the pattern mixture approach, one would estimate the model separately for the six subgroups defined by the number of waves for which data are available. In that case, the quadratic time term could not be estimated for cases with only two waves of data. Clearly, some assumption needs to be made—one can assume that the parameter is equal to the group most similar (those with three waves of data) or the largest group (those with the most data).

The primary disadvantage of this method is that the estimates for any single subgroup are not of special interest. Rather, the estimates of interest are those for the population as a whole, effectively requiring the analyst to combine estimates across the subgroups (Fitzmaurice, Laird and Shneyer, 2001). As a result, this methodology requires an additional step.

To avoid this problem in a longitudinal analysis, Fitzmaurice and colleagues suggest an alternative method for parameterizing the outcome equation predicting Y. In particular, to capture the relationship between the outcome of interest and the likelihood of nonresponse, they suggest including the number of waves an individual participates in a study as a set of covariates, with one dummy variable representing each of the possible number of waves. They demonstrate that one can center those variables in a way that the main effect of the other covariates represents the population estimate of the parameters of interest. The centering involves estimating a multinomial logit model predicting the number of waves an individual participates. One then uses the resulting parameter estimates to calculate the predicted probability of participating in a given number of waves. A function of those predictions is then used to center the participation dummies. For details, see Fitzmaurice, Laird and Shneyer (2001). As a result, the predicted value for Y does not depend on the participation dummies (the expected values of which are zero). As a result, the parameters on the other covariates (such as gender or race) have their standard interpretation.

3 Empirical application: methods

One effort to improve mental health services for children and youth revolves around the philosophy of "a system of care". Key components of this philosophy include community-based alternatives to out-of-home placements, family involvement, cultural sensitivity, and interagency cooperation and coordination (Stroul and Friedman, 1996).

A federal effort to promote system reform according to these principles involves the Comprehensive Community Mental Health Services for Children and Their Families Program, established in 1992 by the Center for Mental Health Services (CMHS) in the Substance Abuse and Mental Health Services Administration (Center for Mental Health Services, 1999). This program has provided over $100 million to 121 communities over the past 14 years for the development of local systems of care and has served over 67,000 children and their families nationwide.

The CMHS grants support expanded provision of community-based, culturally sensitive services and encourage the development of interagency coordination. With regard to the former, recipient communities must develop a full range of services, including diagnosis and evaluation, case management, outpatient therapy, 24-hour emergency services, intensive home-based care, intensive day treatment, respite care, therapeutic foster care, and transition services. To ensure that programs continue after federal funding, local sites are required to develop a match for federal funds; the level of matching funds must increase during the course of the grant.

Within this rather broad framework, local sites can tailor their system to the strengths, resources, and needs of their community. As a result, the sites differ in terms of the age and needs of the children and adolescents served as well as the portals through which children enter the system (e.g., child welfare, juvenile justice, community mental health). They also differ in the services offered and in the settings in which services are delivered (e.g., school or home).

The program is the focus of a national, multisite, multicomponent evaluation, which provides the data for this article. One critical evaluation component consists of a quasi-experimental study that matches and compares funded system-of-care communities with similar nonfunded communities. Our analyses here focus on two of these pairs, one in Nebraska and one in Alabama. The catchment area for the system-of-care grant-funded program in Alabama is Jefferson County (the Jefferson County Community Partnership) and includes the city of Birmingham. The matched comparison community is located in four contiguous counties that are served by the Montgomery Area Mental Health Authority. The implementation of interagency approaches by the Jefferson County Community Partnership includes particular focus on children with mental health or behavioral problems who are involved in the juvenile justice system. In Nebraska, Behavioral Health Region III is the system-of-care program (Nebraska Family Central) and is based in Kearney. In Nebraska Behavioral Health Region IV, the comparison community is based in Norfolk. Each of these regions covers a 22-county rural area with a span of approximately 15,000 square miles.

As part of the evaluation, a sample of 939 children and adolescents aged 4 to 17 with serious emotional and behavioral problems who were using mental health services were recruited for the longitudinal comparison study along with their caregivers. Study enrollment began in August 1999 and continued through May 2003 with follow-up data collection continuing through May 2004. For most youth, entry into the study coincided with entry into services. For youths who had received services in the past, entry into the study coincided with a new episode of care.

Data on the youths' mental health and on family demographics were collected through face-to-face interviews with caregivers and their children. Interviews were conducted at study entry and then at subsequent six-month intervals. The caregiver interviews provided a wide range of information on factors related to the need for mental health services. These characteristics included demographic information on the children, Medicaid enrollment status, and child and family risk factors, and provided the basis for matching across sites. The interviews also provided the outcomes for our analyses.

Study participants were generally between the ages of 4 and 17 at recruitment. The median age was fairly comparable across the four groups defined by site and SOC status, ranging from 12 to 14. About one-third of participants were female and ranged from 27% (NE SOC) to 44% (AL TAU). As one would expect, the racial and ethnic composition of the groups differed substantially between AL and NE. 71% and 83% of the NE TAU and SOC groups, respectively, were non-Latino white. Those figures were 31% and 35% in Alabama. In NE, the non-whites were predominantly Latino, with some African-Americans and Native Americans. In AL, the majority of the sample was African-American.

These outcomes included well-accepted measures of the child's mental health, such as the Child and Adolescent Functional Assessment Scale (CAFAS; Hodges, 1990; Hodges and Gust, 1995; Hodges and Wong, 1997) and the Child Behavior Checklist (CBCL; Achenbach, 1978, 1991; Achenbach and Edelbrock, 1979, 1981). The CAFAS assesses child functioning in eight domains, while the CBCL measures behavioral and emotional symptoms. Higher CBCL scores indicate more behavioral problems. A higher CAFAS score indicates more functional impairment. Reductions in either measure represents improvements in the child's mental health.

Data are available for as many as seven waves of data collection for participating youth.

4 Empirical applications: results

4.1 Descriptive statistics

Table 12.1 provides descriptive statistics for the outcome measures. Several patterns are apparent. First, one can see that over time attrition is fairly high. By the seventh wave, at least three-fourths of the sample had attrited at each site. At both sites and regardless of SOC status, one-third of the sample was lost by the fourth wave. Second, attrition varies with SOC status, but the nature of that difference varied across sites. In Alabama, attrition was higher in the SOC site; the pattern was reversed in Nebraska. Third, regardless of site or SOC status, children showed improvement over time.

4.2 Growth curves

Tables 12.2 and 12.3 present mixture model estimates of parameters and standard errors using the comparison study data from Alabama and Nebraska.

Model 1 is estimated assuming that dropout is ignorable and includes a set of basic predictors. SOC is an indicator that equals unity if a child is served in a system-of-care community. Time denotes a wave of data collection, and SOC X Time is the interaction between SOC and Time indicators. Model 2 assumes that dropout is nonignorable, and following Fitzmaurice et al., introduces additional controls that include the residuals from the first-stage multinomial logit and the interactions of the sum of the first-stage residuals with the basic predictors described above.[3]

[3] The models presented here were estimated assuming all variances and covariances to be distinct. We have re-estimated the models allowing a distinct variance for each random effect within a random-effects equation and restricting all covariances to be zero. The conclusions were not affected by the specification of the variance-covariance matrix. The additional results are available upon request.

Table 12.1 Descriptive statistics, by wave and site

		Alabama Pair									
		System of Care					*Comparison*				
		CBCL				*%Missing*	*CBCL*				*%Missing*
Wave		*Obs*	*Ext*	*Int*	*CAFAS*		*Obs*	*Ext*	*Int*	*CAFAS*	
1	mean	202	69.11	61.58	3.93	0%	189	66.68	63.67	3.49	0%
	sd		10.61	10.93	1.01			11.70	11.85	1.03	
2	mean	146	66.27	59.85	3.52	28%	167	64.60	61.31	3.23	12%
	sd		11.95	11.50	1.13			11.69	12.60	1.11	
3	mean	111	64.03	56.19	3.21	45%	153	62.76	58.54	3.10	19%
	sd		10.73	12.28	1.19			11.92	12.99	1.09	
4	mean	80	63.99	58.06	3.21	60%	124	60.90	57.47	3.10	34%
	sd		13.94	12.59	1.32			12.07	12.92	0.98	
5	mean	63	63.71	57.38	2.95	69%	101	60.13	56.33	3.12	47%
	sd		10.84	11.16	1.11			10.47	11.39	1.08	
6	mean	47	62.55	57.06	3.00	77%	78	56.65	53.78	3.00	59%
	sd		11.78	12.69	1.13			12.41	11.82	1.10	
7	mean	20	61.35	55.05	3.00	90%	48	56.79	51.10	2.92	75%
	sd		7.01	9.72	1.08			11.02	10.90	1.03	
		Nebraska Pair									
		System of Care					*Comparison*				
1	mean	321	70.14	66.24	4.00	0%	222	66.83	62.73	3.92	2%
	sd		9.41	10.80	0.87			10.24	10.16	1.03	
2	mean	286	65.35	61.79	3.53	11%	198	62.27	58.87	3.40	12%
	sd		11.05	11.68	1.03			11.25	11.78	1.17	
3	mean	242	62.43	58.55	3.12	25%	163	59.82	57.16	3.08	28%
	sd		10.79	11.19	1.04			12.17	11.18	1.25	
4	mean	192	61.67	58.29	3.14	40%	118	59.44	56.08	3.02	48%
	sd		11.46	11.96	1.05			12.30	11.77	1.21	
5	mean	144	60.40	56.13	2.90	55%	82	59.11	54.87	3.01	64%
	sd		11.66	12.32	1.16			9.89	11.07	1.20	
6	mean	109	59.50	55.39	2.84	66%	48	59.27	54.98	2.92	79%
	sd		11.27	12.40	1.08			11.33	12.79	1.11	
7	mean	88	58.35	53.43	2.84	73%	29	55.66	50.07	2.50	87%
	sd		12.41	13.66	1.19			9.21	11.78	1.17	

Table 12.2 Alabama comparison study: estimates of treatment effects

ALABAMA						
	CBCL Internalizing		*CBCL Externalizing*		*CAFAS*	
Parameter	*M1*	*M2*	*M1*	*M2*	*M1*	*M2*
Intercept	65.10*** (0.90)	64.82*** (0.98)	67.77*** (0.87)	67.53*** (0.93)	3.41*** (0.08)	3.48*** (0.09)
SOC	−3.18** (1.29)	−2.74** (1.39)	1.95 (1.23)	2.36* (1.33)	0.63*** (0.12)	0.55*** (0.12)
Time	−1.83*** (0.21)	−1.72*** (0.26)	−1.45*** (0.18)	−1.22*** (0.23)	−0.04** (0.02)	−0.03 (0.02)
SOC X Time	0.66* (0.34)	0.42 (0.38)	−0.09 (0.29)	−0.43 (0.33)	−0.16*** (0.03)	−0.17*** (0.04)
$D^*{}_2$		−0.48 (2.42)		0.55 (2.34)		0.92*** (0.21)
$D^*{}_3$		−0.37 (2.48)		2.55 (2.43)		0.61*** (0.22)
$D^*{}_4$		1.52 (2.28)		1.97 (2.23)		0.68*** (0.20)
$D^*{}_5$		−1.21 (2.40)		0.49 (2.37)		0.58*** (0.21)
$D^*{}_6$		1.27 (2.42)		2.24 (2.39)		0.72*** (0.21)
$D^*{}_7$		1.20 (2.31)		1.17 (2.26)		0.49** (0.20)
D^*X SOC		1.45 (3.46)		−1.02 (3.35)		−0.43 (0.31)
D^*X SOC X Time		−0.82 (0.72)		−1.06* (0.61)		−0.06 (0.07)
D^*X Time		0.32 (0.44)		0.52 (0.37)		−0.02 (0.04)

Note: standard errors are in parentheses.
* significant at 10% level
** significant at 5% level
*** significant at 1% level

Table 12.3 Nebraska comparison study: estimates of treatment effects

NEBRASKA						
	CBCL Internalizing		*CBCL Externalizing*		*CAFAS*	
Parameter	*M1*	*M2*	*M1*	*M2*	*M1*	*M2*
Intercept	64.00*** (0.78)	64.09*** (0.81)	67.59*** (0.75)	67.76*** (0.77)	3.96*** (0.07)	3.96*** (0.08)
SOC	2.96*** (1.01)	2.91*** (1.06)	3.03*** (0.96)	3.23*** (1.01)	0.09 (0.09)	0.14 (0.10)
Time	−2.09*** (0.22)	−2.15*** (0.23)	−2.09*** (0.21)	−2.22*** (0.22)	−0.22*** (0.02)	−0.23*** (0.02)
SOC X Time	−0.11 (0.27)	−0.10 (0.31)	−0.03 (0.27)	−0.09 (0.30)	0.00 (0.03)	−0.02 (0.03)
D*$_2$		1.30 (2.54)		4.57* (2.41)		0.75*** (0.23)
D*$_3$		1.37 (2.36)		4.59** (2.25)		0.73*** (0.21)
D*$_4$		−0.16 (2.31)		1.41 (2.19)		0.56*** (0.21)
D*$_5$		0.61 (2.35)		2.94 (2.23)		0.51** (0.21)
D*$_6$		2.63 (2.39)		5.22** (2.27)		0.70*** (0.21)
D*$_7$		0.64 (2.57)		3.54 (2.45)		0.64*** (0.23)
D*X SOC		−1.74 (2.56)		−2.53 (2.43)		−0.53** (0.23)
D*X SOC X Time		0.31 (0.57)		0.62 (0.56)		0.06 (0.06)
D*X Time		−0.40 (0.47)		−1.15** (0.45)		−0.13*** (0.05)

Note: standard errors are in parentheses.
* significant at 10% level
** significant at 5% level
*** significant at 1% level

The results indicate that accounting for nonrandom dropout rates does have a slight effect on the estimates but does not result in substantive changes in the conclusions. Analysis of data from Alabama sites ($n = 391$) revealed that controlling for nonrandom dropout rates reduces the estimates of the coefficients associated with Time for all three outcomes. When the dependent variable was a CAFAS score, the time effect that had been marginally significant in the basic model, became insignificant. The Model 1 estimates of treatment effects (SOC X Time) indicated that CBCL Internalizing Problems scores improved faster in the control community. Accounting for nonrandom dropout rates reduced the value of the treatment effect and resulted in no statistically significant differences between the two communities. Though no significant treatment effects on CBCL externalizing scores were detected, the magnitude of a coefficient increased substantially (from −0.09 to −0.43) under Model 2. With CAFAS score as a dependent variable, the estimates of both models indicated that children served in a system-of-care improved significantly faster than children served in a more traditional setting.

In contrast to the findings for the Alabama sites, analysis of data from Nebraska ($n = 543$) revealed that controlling for nonrandom dropout rates increases the estimates of the coefficients associated with Time for all three outcomes. Both Model 1 and Model 2 estimates of treatment effects showed no significant differences between the sites in rates of improvement in either of the three outcomes.

5 Discussion

Nonignorable nonresponse remains a challenge for methodologists and applied researchers alike. Fortunately, the range of potential remedies continues to expand in light of both theoretical and computational advances. As discussed, however, no method is likely to produce the “final” or “best” answer in the near term. Rather, the alternative methods represent a range of plausible solutions, and researchers can only rely on their judgement to select an overall strategy (e.g., pattern mixture vs selection) or a specification of a strategy (e.g., which interaction terms to include in the pattern mixture model).

The best approach, therefore, likely involves estimating key model parameters under alternative sets of assumptions. When the results of the analysis are invariant to the handling of missing data, as seems to be the case here, the analyst is left with a rather tidy situation. More challenging are situations where the different methods produce different estimates of key parameters. In that case, the analyst is left to pick among the various models according to which set of assumptions are most tolerable.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist and 1991 Profile*. Burlington, VA: University of Vermont.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics,* 43(1): 49–93.

Fitzmaurice, G. M., Laird, N. M. and Shneyer, L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Statistics in Medicine,* 20(7): 1009–1021.

Foster, E. M. and Bickman, L. (1996). An evaluator’s guide to detecting attrition problems. *Evaluation Review,* 20(6): 695–723.

Foster, E. M., Fang, G. Y. and Conduct Problems Prevention Research Group. (2004). Alternative methods for handling attrition in prevention research: An illustration using data from fast track. *Evaluation Review*, 28: 434–464.

Hodges, K. (1990). *Child and Adolescent Functional Assessment Scale (CAFAS).* Ypsilanti, MI: Eastern Michigan University, Department of Psychology.

Hodges, K. and Gust, J. (1995). Measures of impairment for children and adolescents. *Journal of Mental Health Administration,* 22(4): 403–413.

Kenward, M. G. (1998). Selection models for repeated measurements with nonrandom dropout: An illustration of sensitivity. *Statistics in Medicine,* 17: 2723–2732.

Lillard, L. and Panis, C. (1998). Panel attrition from the Panel Study of Income Dynamics. *Journal of Human Resources,* 33(2): 437–457.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edn. Hoboken, New Jersey: Wiley.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8: 3–15.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods,* 7(2): 147–177.

Verbeek, M. and Nijman, T. (1992). Testing for selectivity bias in panel data models. *International Economic Review,* 33(3): 681–703.

Part III

Descriptive and Causal Analysis in Longitudinal Research

This page intentionally left blank

Chapter 13

Graphical techniques for exploratory and confirmatory analyses of longitudinal data

Garrett M. Fitzmaurice

1 Introduction

The very first steps in an analysis of longitudinal data usually include an examination of simple descriptive statistics, with the goal of obtaining some insights about the patterns of change in the response over time. Although these descriptive statistics can be presented in a table, for comparative purposes a graphical display of the same information is usually far more revealing. Graphical displays are extraordinarily useful techniques for conveying information about the most salient features of longitudinal data. These graphical tools can provide insights about patterns of change in the mean response over time (e.g., linearity or the lack thereof) and the choice of suitable functional forms for covariates. Ordinarily, a graphical assessment of longitudinal data precedes any formal statistical analyses. This preliminary aspect of longitudinal data analysis is, for the most part, exploratory in nature. Graphical techniques also play an important role in the concluding stages of longitudinal data analysis. A final statistical analysis of longitudinal data is not complete without an assessment of the adequacy of the fitted model; the latter often involves a graphical examination of residuals. Plots of residuals are especially helpful for model checking in the confirmatory stages of the analysis. Plots of the residuals are useful not only for revealing systematic trends but also for highlighting anomalies (e.g., potential outliers). In this chapter, we will focus on some graphical techniques commonly used for these two important and complementary aspects of longitudinal data analysis. However, before discussing any particular technique, we introduce two examples that will be used to illustrate the application of these graphical methods. The first example is from a randomized longitudinal clinical trial, the second is from an observational study.

1.1 Treatment of Lead-Exposed Children (TLC) Trial

It is now well-established that exposure to lead can produce cognitive impairment, especially among young children and infants. Although the use of lead as an additive in gasoline has been discontinued, at least in the United States, resulting in a dramatic reduction in airborne lead levels, a small percentage of children continue to be exposed to lead at levels that can produce impairment. Much of this exposure is due to chipping and peeling lead-based paint in older homes. Lead paint chips and lead-contaminated paint dust is ingested by young children during normal teething and hand-to-mouth behavior. The United States Centers for Disease Control and Prevention (CDC) has concluded that children with blood lead levels above 10 micrograms per deciliter (μg/dL) of whole blood are at risk of adverse health effects.

Fortunately, lead poisoning in children is treatable in the sense that there are medical interventions, known as chelation treatments, that can help a child to excrete the lead that has been ingested. Until recently, chelation treatment of children with high levels of blood lead was administered by injection and required hospitalization. A new chelating agent, succimer, enhances urinary excretion of lead and has the distinct advantage that it can be given orally, rather than by injection. In the 1990s, the *Treatment of Lead-Exposed Children (TLC) Trial Group* conducted a placebo-controlled, randomized trial of succimer in children with confirmed blood lead levels of 20–44 μg/dL; levels well above the CDC's threshold for concern about the adverse health effects of exposure to lead (Treatment of Lead-Exposed Children (TLC) Trial Group, 2000; Rogan et al., 2001). The children were aged 12–33 months at enrollment and lived in deteriorating inner-city housing. The mean age of the children at randomization was 2 years and their mean blood lead level was 26 μg/dL. Children received up to three 26-day courses of succimer or placebo and were followed for 3 years. We will focus on longitudinal data on blood lead levels measured at baseline, week 1, week 4, and week 6 on a subset of 100 children from this study who were randomized to placebo (control) or succimer (active treatment).

1.2 MIT Growth and Development Study

The second illustrative example is from a prospective longitudinal study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study (Bandini *et al.*, 2002; Phillips *et al.*, 2003). At the start of the study, all of the girls were premenarcheal and nonobese, as determined by a triceps skinfold thickness less than the 85th percentile. All girls were followed over time according to a schedule of annual measurements until four years after menarche. The final measurement was scheduled on the fourth anniversary of their reported date of menarche. At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis. One of the goals of the study was to examine changes in body fat accretion before and after menarche.

For the purposes of analyses, the "time" of measurement is calibrated as "time since menarche"; therefore it can be positive (for measurements after the reported date of menarche) or negative (for measurements prior to menarche). Thus, although the measurement protocol is the same for all girls if the timing of measurement is defined as the time since the baseline measurement, it is highly irregular when the timing of measurements is defined as the time since a girl experienced menarche. Relative to menarche, each girl is measured at a unique set of occasions, with few observation times coinciding. In this data set there are a total of 1049 individual percent body fat measurements, with an average of 3.1 measurements during the premenarcheal period and 3.5 measurements during the postmenarcheal period.

2 Graphical exploration of longitudinal data

The formal statistical analysis of longitudinal data should always be preceded by simple graphical displays of the data. A natural way to display longitudinal data is through the use of a standard scatter-plot, with the responses on the vertical axis and the measurement times on the horizontal axis. We refer to such a plot as a *time plot*. Unfortunately, a time plot of longitudinal data may not always be very helpful or readily interpretable. In many longitudinal studies there are a fixed set of common measurements occasions for all study participants; we refer to this as a *balanced* design. For example, the TLC trial discussed earlier has a balanced design, with all children measured at the same set of occasions: baseline (or week 0), week 1, week 4, and week 6. In a balanced design, a time plot of the raw data results in many overlapping data points at each measurement occasion. This can make it difficult to determine any trends in the mean response over time. In addition, such a plot does not indicate which data points represent repeated measurements on the same individual. To circumvent the latter problem, the time plot can be supplemented by connecting successive repeated measures on the same individual with straight lines. However, the resulting line segments do not necessarily enhance the time plot; indeed, more often than not, it can result in a "spaghetti" plot that is not very informative about overall trends in the response over time.

Some of the aforementioned problems with the time plot of longitudinal data can be illustrated using data from the *Treatment of Lead-Exposed Children Trial*. Figure 13.1 displays a time plot of the blood lead level data for the group of children randomized to succimer. Because the data points overlap at the common set of four measurement occasions, it is difficult to discern any pattern in the mean response trend over time. Perhaps the only useful source of information provided by this simple time plot of the raw data concerns the presence of outliers in the data and whether the variability in the data changes discernibly with time. For example, there appears to be an outlying observation at week 6, corresponding to a blood lead level of 64 μg/dL.

In Figure 13.2 the time plot of blood lead levels is supplemented with line segments joining successive measures on the same individual. Figure 13.2 is only marginally more informative about trends in the mean response over time than Figure 13.1. The appeal of joining line segments is that it allows us to distinguish which data points represent repeated measurements on the same individual. However, it can be very difficult to track the response profile of any particular individual when the plot contains longitudinal data on many individuals. With too much "spaghetti", the information conveyed by this plot is difficult to digest. As a result, it may be more useful to present the time plot with joined line segments for only a relatively small random sample of the study participants.

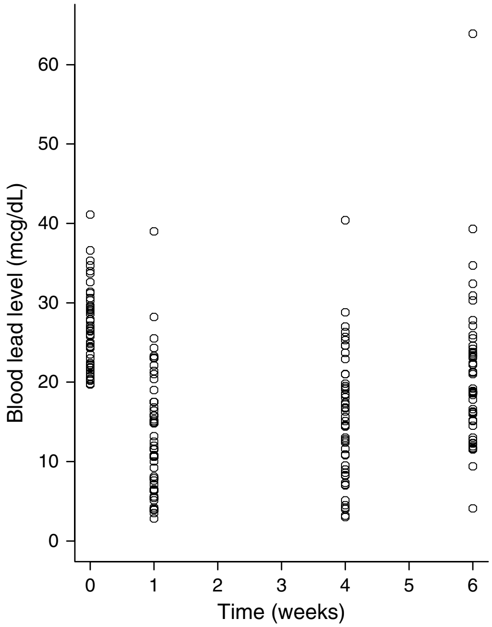


Figure 13.1 Time plot of blood lead levels at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial

Because of the aforementioned problems with time plots of the raw data, it is usually more informative to display a time plot of the mean response, with successive means joined by straight lines. In addition, time plots of the mean response for different levels of discrete covariates (e.g., different intervention or

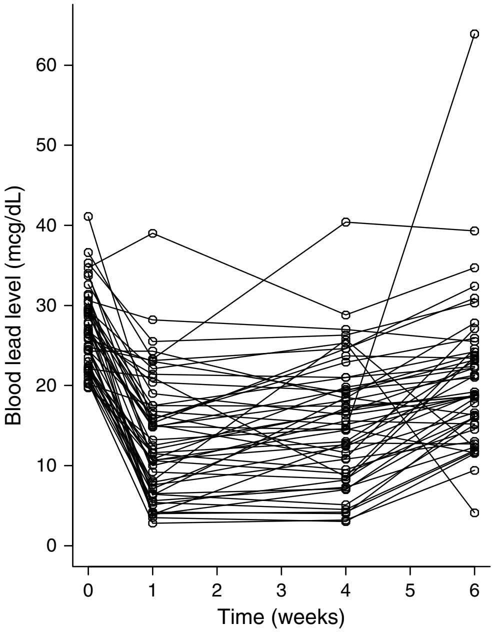


Figure 13.2 Time plot, with joined line segments, of blood lead levels at baseline (week 0), week 1, week 4, and week 6 for children from the succimer group in the TLC trial

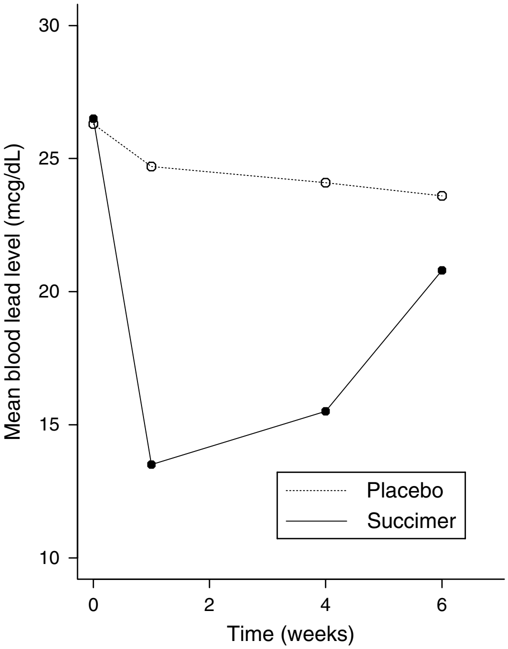


Figure 13.3 Time plot of the mean blood lead levels at baseline (week 0), week 1, week 4, and week 6 in the succimer and placebo groups

treatment groups) can be overlayed on the same graph. The construction of such a plot is relatively straightforward when the timing of the repeated measures is the same for all individuals. The time plots can also be enhanced by including standard error bars for the mean response at each occasion. For example, Figure 13.3 displays the mean blood lead levels in the succimer and placebo groups at weeks 0, 1, 4, and 6. From this simple display it is readily apparent that the effect of succimer is greater after one week of treatment and that there appears to be a rebound effect thereafter. Overall, a graphical display of the mean response can be quite enlightening and can provide the basis for choosing an appropriate model for the analysis of change over time. For example, the time plot of the mean response in Figure 13.3 suggests that the analysis of the blood lead levels at all four occasions may require nonlinear (e.g., quadratic) or perhaps piecewise linear trends over time.

The construction of time plots of the mean response is less straightforward when a covariate of interest is quantitative (e.g., dose of drug). For the purposes of producing a graphical display of the mean response trend, one simple, but often quite effective, approach is to construct a small number of groupings or “reference categories” for the quantitative covariate in question. For ease of exposition, we consider three groupings of the quantitative covariate that can be denoted as “low”, “medium”, and “high”. Given this set of reference categories, the construction of the time plot of the mean response trend can proceed along exactly the same lines as for the case of a truly discrete covariate having only three levels. That is, we can simply plot the mean response trends overlayed for the different values of the reference categories. However, it must be acknowledged that the number and choices of reference groups are, to some extent, arbitrary.

So far, our discussion has assumed that many, if not all, individuals are measured at the same set of occasions. When the times of measurement are not the same, construction of time plots of the mean response can pose difficulties due to sparseness of data at any particular occasion. For example, Figure 13.4 displays a time plot (time relative to age of menarche), with joined line segments, of longitudinal data on percent body fat in the cohort of 162 girls from the MIT Growth and Development Study (Bandini et al., 2002; Phillips et al., 2003). Here, because each girl is measured at a unique set of occasions, with few observation times coinciding, construction of a plot of the mean response over time is difficult due to sparseness of data at any particular time. For example, it is difficult to precisely estimate the mean percent body fat 2 years after menarche because there are so few observations at that particular time. Moreover, it is difficult to discern whether the changes in percent body fat in the premenarcheal period are similar to the changes in the postmenarcheal period from this "spaghetti" plot.

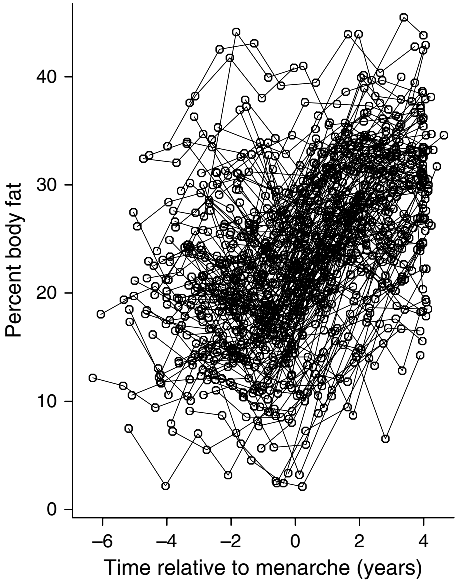


Figure 13.4 Time plot of percent body fat against time, relative to age of menarche (in years)

In cases where the design is highly unbalanced (i.e., repeated measurements are not obtained at a common set of occasions), it is helpful to produce a "smoothed" plot of the mean response trend over time. A smooth plot of the trend can be obtained using a variety of "smoothing techniques". Many of these smoothing techniques approach the estimation of the mean response at any distinct time by considering not only the observations at that occasion but also "neighboring" observations. That is, the estimated mean is based on observations taken before, at, and after the time of interest. Typically, the mean response at any time, say t, is taken to be a weighted average of the observations in some close proximity or neighborhood of time t.

One popular smoothing technique is locally weighted regression or *lowess* (Cleveland, 1979). The lowess estimate of the mean response at time t is determined by fitting a straight line to the observations that fall within a "window" centered at time t. The fitted regression line is obtained using a robust regression technique that gives more weight to observations close to the center of the window and that also down-weights potential outliers. The entire lowess curve is obtained by moving a window of fixed width from the first measurement occasion to the last, and repeating the process at every time. Figure 13.5 displays a lowess curve for the percent body fat data described earlier. Unlike the time plot of the raw data in Figure 13.4, the lowess curve is informative about changes in percent body fat before and after menarche. The smooth curve produced by the lowess procedure reveals that the mean response increases gently during the premenarcheal period and then rises steeply during the postmenarcheal period.

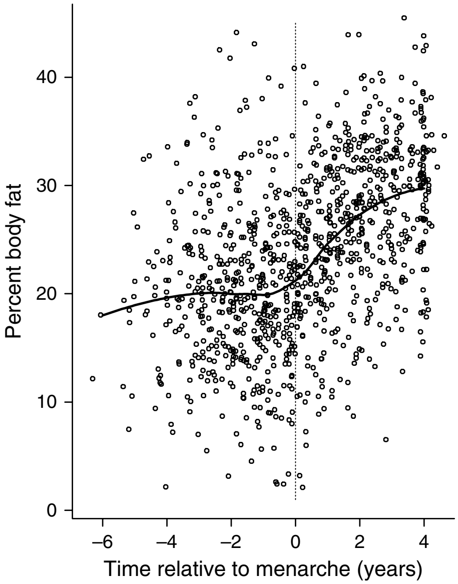


Figure 13.5 Time plot of percent body fat against time, relative to age of menarche (in years), with *lowess* smoothed curve

In summary, graphical techniques have an important role in the early stages of longitudinal data analysis. Although time plots of the raw data, with or without joined line segments, can be difficult to interpret, especially when the number of observations is relatively large, time plots of the mean response can be very informative. Time plots of the mean response are easy to construct when the study design is balanced over time; for highly unbalanced data, various smoothing techniques can be used. A time plot of the mean response can provide useful insights about the general patterns of change over time and possible functional forms for the covariates. For example, Figure 13.3 suggests that the analysis of the blood lead levels at all four occasions may require nonlinear trends over time, especially for the succimer group. Similarly, Figure 13.5 suggests that the growth rate for percent body fat prior to menarche is relatively flat and might be well-approximated by a linear trend; however, after menarche, the growth rate increases steeply. Thus, any model for change in percent body fat over time will need to incorporate different trajectories in the pre- and postmenarcheal periods.

3 Graphical model-checking based on residuals

Next, we consider methods for assessing the adequacy of models for longitudinal data. Typically, the analysis of longitudinal data focuses on changes in the mean response over time, and on the relation of these changes to covariates. For that reason, we concentrate on residual diagnostics for assessing the adequacy of the model for the mean response. Methods of assessing the model for the covariance are mentioned only briefly at the end of this chapter; readers interested in the latter topic are directed to Chapter 9 of Fitzmaurice, Laird and Ware (2004). Also, for ease of exposition, we focus on the assessment of models for longitudinal data where the response variable is continuous; similar techniques can be applied when the response is binary, ordinal or count data.

Methods for residual analyses are well developed for standard regression settings with independent observations on a univariate response; see Cook and Weisberg (1982) for a comprehensive description of techniques for residual analysis. In principle, many of the same properties of residual analysis can be extended to the longitudinal setting, with relatively minor modifications. In this section we also consider some recently developed techniques, based on aggregating residuals, that put residual diagnostics on a somewhat more objective footing.

3.1 Raw residuals

Before we begin our discussion of residuals we must introduce some notation. We assume that N subjects are measured repeatedly over time. We let Y_{ij} denote the response variable for the i^{th} subject on the j^{th} measurement occasion.

In principle, the response variable could be continuous, binary, or a count; however, for ease of exposition, we focus on the case where Y_{ij} is continuous. To accommodate unbalanced data, we assume that there are n_i repeated measurements of the response on the i^{th} subject and that each Y_{ij} is observed at time t_{ij}. The response variables for the i^{th} subject can be grouped into an $n_i \times 1$ vector

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \quad i = 1, \ldots, N$$

where the vectors of responses, Y_i, are assumed to be independent of one another (but the repeated measures on the same subject are emphatically not assumed to be independent). Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \quad i = 1, \ldots, N; \ j = 1, \ldots, n_i$$

The vector of covariates at the n_i occasions can be grouped into a $n_i \times p$ matrix denoted by X_i.

We assume the following linear model for the vector of continuous responses, Y_i,

$$Y_i = X_i\beta + e_i \qquad (1)$$

where the unknown regression parameters can be grouped together into a $p \times 1$ vector, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $e_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})'$ is an $n_i \times 1$ vector of random errors. The random errors, e_{ij}, have mean zero and represent deviations of the responses from their corresponding predicted means

$$E(Y_{ij}|X_{ij}) = X'_{ij}\beta = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} \qquad (2)$$

Typically, although not always, $X_{ij1} = 1$ for all i and j, and then β_1 is the intercept term in the model.

Thus far we have made no distributional assumptions about Y_i. The only assumption made is that the mean of the longitudinal response vector is related to the covariates via the linear regression model given above. When Y_i is a vector of continuous response, it is commonly assumed that it has a multivariate normal distribution, with mean response vector

$$E(Y_i) = \mu_i = X_i\beta$$

and covariance matrix,

$$\Sigma_i = \mathrm{Cov}(Y_i)$$

Recall that the multivariate normal distribution is completely specified by the vector of means, μ_i, and the covariance matrix, Σ_i. The covariance can be modelled directly or via the introduction of random effects (e.g., linear mixed effects models). This completes our specification of the model for Y_i.

Given a regression model for the mean response, specified by equation (1), we can define a vector of residuals for each individual,

$$r_i = Y_i - X_i\widehat{\beta} \qquad (3)$$

The vector of residuals has mean zero and provides an estimate of the vector of errors,

$$e_i = Y_i - X_i\beta$$

These residuals can be used to check for any systematic departures from the regression model for the mean response. For example, a scatter-plot of the residuals

$$r_{ij} = Y_{ij} - X'_{ij}\widehat{\beta}$$

against the predicted mean response

$$\widehat{\mu}_{ij} = X'_{ij}\widehat{\beta}$$

can be examined for the appearance of any systematic trend. The fitting of a smooth curve (e.g., a *lowess* curve) to the scatter-plot can often help in judging whether curvature is present. In a correctly specified model for the mean response, the plot should display no systematic pattern, with a more or less random scatter around a constant mean of zero. Similarly, scatter-plots of the residuals against selected covariates from the model for the mean can be examined for any systematic trends. Such a trend may indicate the omission of a quadratic term or the need for transformation of the covariate.

For most practical purposes, graphical displays of the residuals can be used to detect discrepancies in the model for the mean response or the presence of outlying observations that require further investigation. However, there are two properties of the residuals from an analysis of longitudinal data that set them apart from residuals in a standard regression with independent observations on a univariate response. First, the components of the vector of residuals,

$$r_i = Y_i - X_i\widehat{\beta}$$

are correlated and do not necessarily have constant variance. Because the residuals have approximate covariance matrix, $\text{Cov}(r_i) \approx \text{Cov}(e_i) = \Sigma_i$, this has important implications for the examination of plots of the residuals. First, standard residual diagnostics for examining either the homogeneity of the residual variance or autocorrelation among the residuals should be avoided altogether. Second, although residuals from a univariate linear regression are uncorrelated with the covariates, the residuals from a regression analysis of longitudinal data may be correlated with the covariates. As a result, there may be an apparent systematic trend in the scatter-plot of the residuals against a selected covariate.

3.2 Transformed residuals

To circumvent some of the aforementioned problems, we can transform the residuals so that they have constant variance and zero correlation, thereby mimicking residuals from a standard linear regression. This can be achieved using a well-known technique called the *Cholesky decomposition* (or *Cholesky factorization*). Given an estimate of the approximate covariance matrix for the residuals, $\widehat{\Sigma}_i$, the Cholesky decomposition of $\widehat{\Sigma}_i$ can be used to create a lower triangular matrix, L_i, such that

$$\widehat{\Sigma}_i = L_i L_i'$$

Note that a lower triangular matrix is simply one with all zeros above the diagonal. We can then use the matrix L_i or, more specifically, L_i^{-1}, to take us from a set of correlated residuals with heterogeneous variances to a set of transformed residuals,

$$r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\widehat{\beta}) \qquad (4)$$

that are uncorrelated and have unit variance.

Given the set of transformed residuals, r_i^*, all of the usual residual diagnostics for standard linear regression can be applied. For example, we can construct a scatter-plot of the transformed residuals, r_{ij}^*, versus the transformed predicted values, $\widehat{\mu}_{ij}^*$, where

$$\widehat{\mu}_i^* = L_i^{-1}\,\widehat{\mu}_i = L_i^{-1}\,X_i\widehat{\beta}$$

In a correctly specified model, this plot should display no systematic pattern, with a random scatter around a constant mean of zero and with a constant range for varying $\widehat{\mu}_{ij}^*$. Similarly, we can construct a scatter-plot of the transformed residuals versus selected transformed covariates. With longitudinal data, a scatter-plot of the transformed residuals versus transformed time (or age) can be particularly useful for assessing the adequacy of the model assumptions about patterns of change in the mean response over time. We note that standard linear regression programs can be used to automate the production of residual diagnostics. That is, standard residual diagnostics can be applied after refitting a standard linear regression of Y_i^* on X_i^*, where $Y_i^* = L_i^{-1}\,Y_i$ and $X_i^* = L_i^{-1}\,X_i$. For a more detailed discussion of the generalization of residual diagnostics to longitudinal data, the interested reader is referred to articles by Waternaux et al. (1989) and Waternaux and Ware (1991).

Finally, the transformed residuals also make it somewhat easier to identify outliers, both outlying *observations* and outlying *individuals*. Outlying observations may be indicated by large residuals (e.g., a residual with absolute value greater than 2 or 3). Because we are focusing on the most extreme values of the residuals, the distribution of these extremes is somewhat more complicated than a standard normal distribution. In general, we recommend careful examination of the most extreme residuals while recognizing that extreme residuals will occur with predictable regularity; for example, with 1000 residuals, the expected number of residuals whose absolute value exceeds 2 is approximately $1000 \times 0.05 = 50$. Alternatively, an outlying individual can be identified by first calculating a summary measure of multivariate distance between their observed and fitted responses, based on the *Mahalanobis distance*,

$$d_i = r_i^{*\prime}\, r_i^* \tag{5}$$

If the model is correctly specified, the distance given by equation (5) has an approximate chi-squared distribution with degrees of freedom (df) equal to the dimension of r_i^* (i.e, df $= n_i$, the number of repeated measurements on the i^{th} subject). Outlying individuals will have distances, d_i, that have small associated p-values. The p-values provide a common metric for comparing and detecting large values of d_i, corresponding to unusual or outlying individuals, when the number of repeated measurements varies across subjects. Once again, we caution that the distribution of the extremes is somewhat more complicated and it is important to recognize that extremes will occur with predictable regularity.

To illustrate the use of raw and transformed residuals, we will consider assessing the adequacy of a longitudinal model for the body fat accretion data from the MIT Growth and Development Study. Recall that the data are from a prospective longitudinal study examining changes in body fat before and after menarche in a cohort of 162 girls. For the analysis of these data, "time" was coded as time since age of menarche and could be positive or negative. We consider the hypothesis that percent body fat increases linearly with age, but with different slopes before and after menarche. Specifically, we assume that each girl has a piecewise linear spline growth curve with a knot at the time of menarche and fit the following linear mixed effects model,

$$E(Y_{ij}|b_{1i}, b_{2i}, b_{3i}) = \beta_1 + \beta_2\, t_{ij} + \beta_3\,(t_{ij})_+ \\ + b_{1i} + b_{2i}\, t_{ij} + b_{3i}\,(t_{ij})_+$$

where t_{ij} denotes the time of the j^{th} measurement on the i^{th} subject before or after menarche (i.e., $t_{ij} = 0$ at menarche), $(t_{ij})_+ = t_{ij}$ if $t_{ij} > 0$ and $(t_{ij})_+ = 0$ if $t_{ij} \leq 0$. The random effects, b_{1i}, b_{2i}, b_{3i}, are assumed to have a multivariate normal distribution, with zero mean. (An excellent review of linear mixed effects models with piecewise linear trends can be found in Naumova et al. (2001)). In this model, each girl's growth curve can be described with an intercept and two slopes, one slope for the changes in response before menarche, another slope for the changes in response after menarche.

The restricted maximum likelihood (REML) estimates of the fixed effects are displayed in Table 13.1. Based on the magnitude of the estimate of β_3, relative to its standard error, it can be concluded that there is a significant difference between the slopes before and after menarche. In particular, the estimated premenarcheal slope is rather shallow

Table 13.1 Estimated regression coefficients (fixed effects) and standard errors for the piecewise linear model for the percent body fat data

Variable	*Estimate*	*SE*	*Z*
Intercept	21.3614	0.5646	37.84
Time	0.4171	0.1572	2.65
$(\text{Time})_+$	2.0471	0.2280	8.98

(0.42) and indicates that the annual rate of body fat accretion is less that 0.5%. In contrast, the estimated postmenarcheal slope is 2.46 (2.047 + 0.417) and indicates that the annual rate of body fat accretion is approximately 2.5%, almost six times higher than the corresponding rate in the premenarcheal period.

Next we use residual diagnostics to assess the adequacy of the fitted model. Based on the Cholesky decomposition of the estimated covariance matrix, $\widehat{\Sigma}_i$, we can calculate transformed residuals,

$$r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\widehat{\beta})$$

where $\widehat{\Sigma}_i = L_i L_i'$. For illustrative purposes, we also examine the untransformed residuals and compare the diagnostic plots based on these two types of residuals.

Normal quantile plots of the transformed and untransformed residuals are presented in Figure 13.6 and do not indicate any systematic departures from a straight line. There is no evidence to suggest any discernible skewness and the normal assumption appears to be tenable. The quantile plot of the transformed residuals does reveal one very extreme observation. However, the number of extreme residuals highlighted by Figure 13.6 is not more than what we would expect due to chance, given a total of 1049 observations.

Next we consider scatter-plots of the transformed and untransformed residuals versus the transformed and untransformed predicted

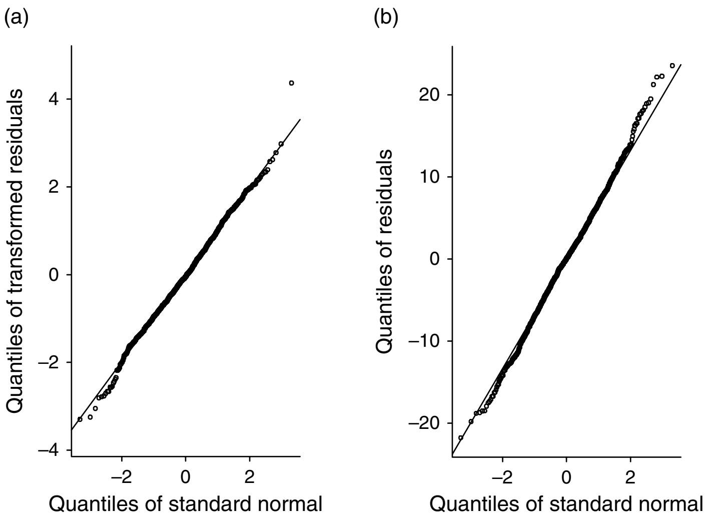


Figure 13.6 Normal quantile plot of (a) the transformed residuals, and (b) the untransformed residuals, for the percent body fat data

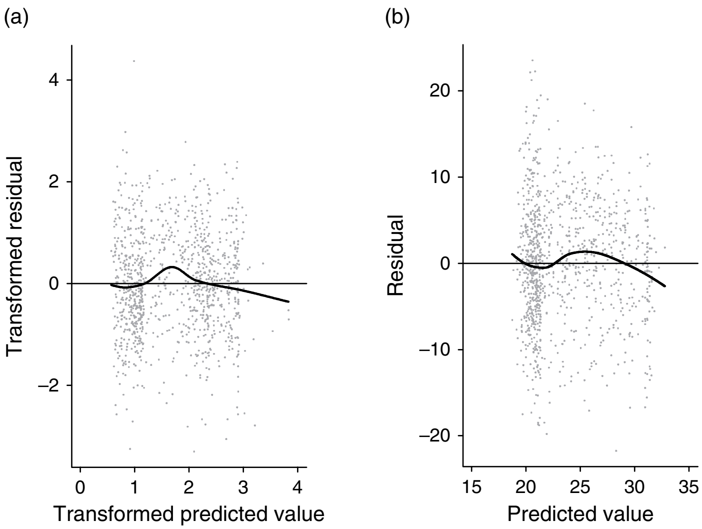


Figure 13.7 Scatter-plot of (a) the transformed residuals versus transformed predicted values, and (b) the untransformed residuals versus predicted values, for the percent body fat data

values respectively. The scatter-plots of the residuals in Figure 13.7 display no obvious systematic pattern, with a random scatter around a constant mean of zero. However, when lowess smoothed curves are superimposed on the scatter-plots, they do reveal some apparent curvature. Focusing on the transformed residuals, there appears to be a quadratic trend, although the fall in the lowess curve at the largest values of the transformed predicted values should be cautiously interpreted as the fitted curve is based on few observations at the extremities and is therefore likely to be unreliable in that region.

Because of the suggestion of curvature in Figure 13.7, we next examine scatter-plots of the (transformed) residuals versus (transformed) time (see Figure 13.8). These scatter-plots of the transformed and untransformed residuals suggest curvature at (untransformed) times corresponding to approximately 2 to 4 years post-menarche. The pattern is more apparent in the scatter-plot of the transformed residuals and can no longer be discounted due to sparseness of the observations at the extremities; here the plots of the transformed and untransformed data give somewhat different impressions. The curvature in the scatter-plots suggests that the model for the mean response might be improved by the inclusion of a quadratic trend in the postmenarcheal period.

Next, we illustrate how the transformed residuals can be used to identify unusual individuals. We can calculate the Mahalanobis distance,

$$d_i = r_i^{*\prime} r_i^*$$

for each girl and then compare the values to reference chi-squared distributions with degrees of freedom (df) equal to the dimension of r_i^* (i.e, df $= n_i$, the number of repeated measurements obtained on each girl). For each girl, we calculated d_i and its associated p-value. There were 7 girls whose d_i yielded p-values less than 0.05 and 2 girls with p-values less than 0.01. Given that the sample is comprised of 162 girls, distances of these magnitudes are to be expected by chance alone.

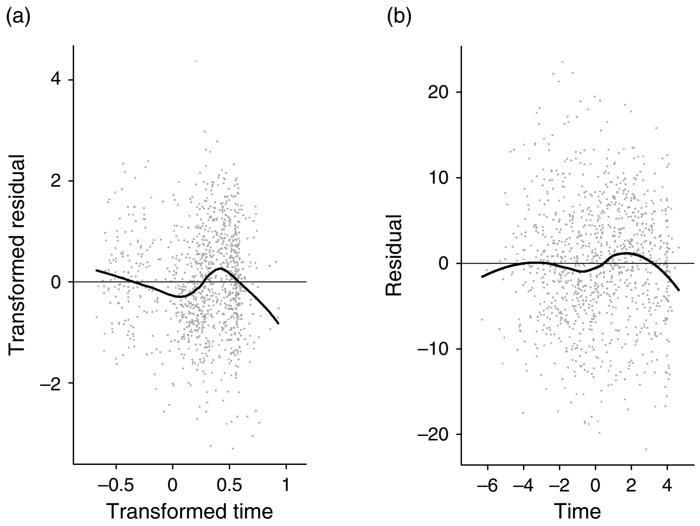


Figure 13.8 Scatter-plot of (a) the transformed residuals versus transformed time, and (b) the untransformed residuals versus time, for the percent body fat data

3.3 Aggregating residuals

So far, much of the discussion of residual diagnostics has focused on graphical techniques for assessing the adequacy of the model for the mean response. With appropriate transformations, we have seen that residual diagnostics developed for standard linear regression can be extended to the longitudinal setting. An acknowledged difficulty with conventional residual diagnostics is that they are somewhat subjective in nature. What appears to be a random scatter to one individual, might be considered evidence of systematic trend to another. That is, it can be very difficult to discern whether an apparent trend in a scatter-plot of the residual reflects some aspect of model misspecification or is simply a reflection of natural variation. McCullagh and Nelder (1989, pp. 392–393) aptly summarize this problem when they state that "the practical problem is that any finite set of residuals can be made to yield some kind of pattern if we look hard enough, so that we have to guard against over-interpretation."

Recently, model-checking techniques based on "cumulative sums" and "moving sums" of residuals have been developed to help discern the "signal" from the "noise". The basic idea is to aggregate the residuals over certain coordinates. The coordinates typically used for these sums of residuals are the individual covariates (e.g., X_{ijk}, the k^{th} covariate) and the fitted values, $X'_{ij}\widehat{\beta}$. The advantage of working with sums of residuals, rather than raw or transformed residuals, is that a reference distribution is available to ascertain their natural variation. That is, we can compare the *observed* sums of the residuals, both graphically and numerically, to a reference distribution under the assumption of a correctly specified model for the mean. This allows us to determine whether any apparent pattern is evidence of a systematic trend or simply due to natural variation. This removes a large degree of subjectivity from the assessment of graphical displays of residuals and places residual diagnostics on a more objective footing.

Recall that the raw residuals are defined as the difference between the observed and fitted values of the response,

$$r_{ij} = Y_{ij} - X'_{ij}\widehat{\beta}$$

If the model for the mean is correctly specified, these residuals are centered at zero. To check the functional form of any covariate, say X_{ijk}, the k^{th} covariate, we can define the cumulative sum of the residuals over values of X_{ijk},

$$W_k(x) = \sqrt{N} \sum_{i=1}^{N} \sum_{j=1}^{n_i} I(X_{ijk} \leq x) r_{ij}$$

where $I(\cdot)$ is the indicator function. For any given x, $W_k(x)$ is the sum of residuals for all values of X_{ijk} less than or equal to x. The process $W_k(x)$ is a step function with possible jumps (either increases or decreases) at all of the distinct values for X_{ijk}. The cumulative sum of residuals can be defined similarly with respect to any other covariate. In addition, we can construct the cumulative sum of residuals over the fitted values, denoted by $W_f(x)$,

$$W_f(x) = \sqrt{N} \sum_{i=1}^{N} \sum_{j=1}^{n_i} I(X_{ij}'\widehat{\beta} \leq x) r_{ij}$$

The cumulative sums, $W_k(x)$, can be used to assess the functional form of the covariates. For example, we can construct a plot of $W_k(x)$ versus x, where for any value of x on the horizonal axis, the corresponding value of $W_k(x)$ on the vertical axis is the cumulative sum of the residuals for covariate values of X_{ijk} less than or equal to x. Evidence of systematic trend in this plot suggests that the functional form of the covariate (e.g., linearity) is not correctly specified and may indicate that a transformation of the covariate or the inclusion of polynomials is required. The cumulative sum, $W_f(x)$, is useful for assessing the assumption of linearity (or, more generally, the link function). Any evidence of systematic trend in this plot might suggest that either a transformation of Y (i.e., a transformation of the response) or of $E(Y|X)$ (i.e., an alternative link function) is necessary.

If the assumed model for the mean response is correct, then the cumulative sums of residuals are centered at zero. Moreover, we can ascertain the natural variation of the cumulative sum. In particular, the distribution of the cumulative sum can be approximated by that of a Gaussian (or normal) process with zero mean whose realizations can be generated by computer simulation. That is, it is relatively straightforward to generate realizations from the distribution of the cumulative sum, under the assumption that the model for the mean is correct; the technical details are omitted here and the interested reader is referred to Lin, Wei and Yang (2002). Thus, in practical terms, the null distribution of $W_k(x)$ (or $W_f(x)$) is approximated through computer simulation of the zero-mean Gaussian process, denoted by $\widehat{W}_k(x)$ (and $\widehat{W}_6(x)$ respectively). Then, to assess whether any apparent trend in the *observed* cumulative sum of residuals reflects systematic trend rather than chance fluctuations, we can superimpose a number of realizations from the appropriate Gaussian process. To the extent that the curves generated from the null distribution tend to be closer to and intersect zero more often than the observed curve, this provides evidence of lack of fit. This assessment can be put on a more formal footing by comparing the maximum absolute value of the observed cumulative sum to a large number of realizations (say 10,000) from the null distribution. By comparing $\max|W(x)|$, the maximum absolute value of the observed cumulative sum, to $\max|\widehat{W}(x)|$ for each realization from the null distribution, a p-value can be constructed based on the proportion of times that $\max|\widehat{W}(x)| \geq \max|W(x)|$; the latter is referred to as a "supremum" test and provides an omnibus test of model adequacy with respect to the relevant coordinate (e.g., a particular covariate or the fitted values). If the p-value is very small (say less than 0.05 or 0.01), then the model fit can be improved.

There is an alternative way to aggregate the residuals by using a "moving sum" rather than a "cumulative sum". We can define a moving sum of residuals, with "window" b, as follows,

$$W_k(x,b) = \sqrt{N}\sum_{i=1}^{N}\sum_{j=1}^{n_i} I(x-b \leq X_{ijk} \leq x)r_{ij}$$

This represents the sum of residuals in blocks of window size, b. Similarly, to assess linearity (or the link function), we can define a moving sum of residuals with respect to the fitted values,

$$W_f(x,b) = \sqrt{N}\sum_{i=1}^{N}\sum_{j=1}^{n_i} I(x-b \leq X'_{ij}\widehat{\beta} \leq x)r_{ij}$$

A potential advantage of using a moving sum of residuals is that the process is less influenced by the residuals associated with small covariate values. One disadvantage of moving sums, however, is that they require a somewhat arbitrary choice of window size, b. Simulation results suggest that the optimal choice of b is approximately the range of the lower half of the covariate values.

To illustrate the use of sums of residuals, we will consider assessing the adequacy of the longitudinal model for percent body fat introduced earlier. Figure 13.9 shows a plot of the observed cumulative sum of the residuals (solid curve), with respect to the covariate time (relative to age of menarche). On the vertical axis is the cumulative sum of residuals; the horizonal axis denotes time (in years). Superimposed on the graph are two realizations (dotted curves) from the null distribution under the assumption that the model for the mean response is correctly specified. These two realizations are computer simulated from the appropriate Gaussian mean-zero process. By comparing the observed cumulative sum to many different realizations under the null, it is possible to determine whether any apparent trend is systematic or due to chance fluctuations. From Figure 13.9, the simulated realizations produce curves that appear to be closer to

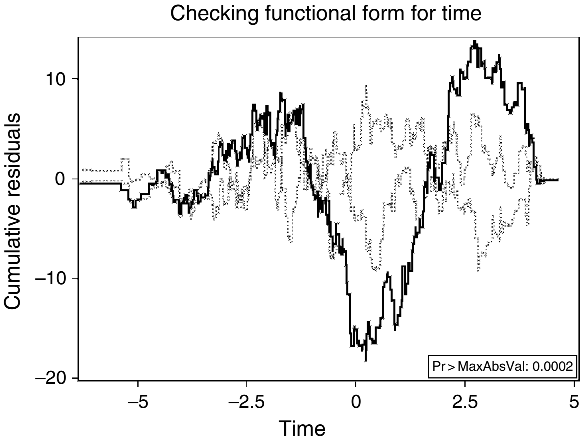


Figure 13.9 Plot of observed cumulative sum of residuals versus time (relative to age of menarche) and 2 simulated realizations from the null distribution assuming a correctly specified model for mean percent body fat. **Note:** Supremum p-value is based on 10,000 simulated realizations from the null distribution

and intersect zero more often that the observed curve. By generating many more such realizations from the null distribution, it is possible to get both a graphical and numerical indication of whether the curve describing the observed cumulative sum displays a systematic pattern or simply natural variation. Figure 13.10 shows a plot of the observed cumulative sum of the residuals and 10,000 realizations from the null distribution. It would appear that the observed cumulative sum displays a systematic pattern. In particular, the observed cumulative sum is too small in the 12 months after menarche (years 0 to 1) and too large 2 to 4 years after menarche. This suggests that the assumed functional form for time, in particular after menarche, may not be adequate. This graphical assessment of fit can be complemented by a numerical assessment. The maximum absolute value of the observed cumulative sum is 18.28. The so-called supremum test yields a p-value of 0.0002, based on the 10,000 simulated

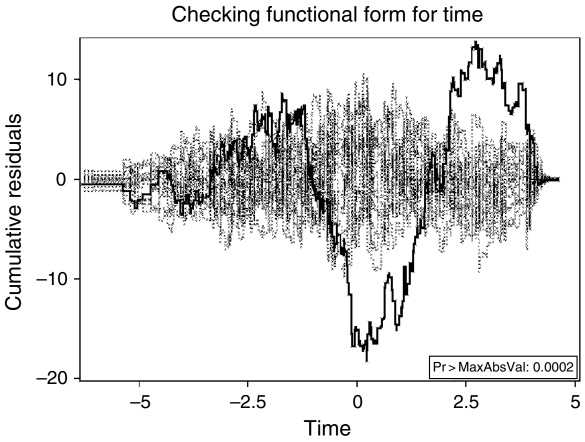


Figure 13.10 Plot of observed cumulative sum of residuals versus time (relative to age of menarche) and 10,000 simulated realizations from the null distribution assuming a correctly specified model for mean percent body fat

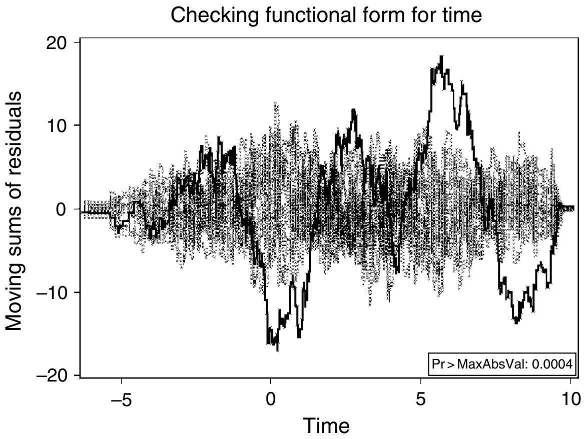


Figure 13.11 Plot of observed moving sum of residuals versus time (relative to age of menarche) and 10,000 simulated realizations from the null distribution assuming a correctly specified model for mean percent body fat

realizations of the process under the null. That is, out of 10,000 simulated realizations, only 2 had a maximum absolute value that exceeded 18.28. Thus, both the graphical and numerical results suggest that the functional form for time, in particular after menarche (time = 0), may be inappropriate.

A similar plot can be constructed based on a moving sum rather than a cumulative sum. Figure 13.11 shows a plot of the observed moving sum of the residuals, with block size equal to half the range of time (approximately 5.5 years). The observed curve in Figure 13.11 also suggests that the moving sum of the residuals is too small in years 0–1 and too large in later years. In a similar fashion, we can complement this graphical display with a numerical assessment. The supremum test yields a p-value equal to 0.0004 (based on 10,000 simulated realizations), suggesting that the functional form for time may be inappropriate.

Next, we consider a refinement to the model for percent body fat to allow for a quadratic trend in the postmenarcheal period. In particular, we assume that each girl has a piecewise linear-quadratic growth curve with a knot at the time of menarche and fit the following linear mixed effects model

$$\begin{aligned} E(Y_{ij}|b_{1i}, b_{2i}, b_{3i}, b_{4i}) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ \\ + \beta_4 (t_{ij})_+^2 + b_{1i} + b_{2i} t_{ij} \\ + b_{3i}(t_{ij})_+ + b_{4i}(t_{ij})_+^2 \end{aligned}$$

where $(t_{ij})_+^2 = t_{ij}^2$ if $t_{ij} > 0$ and $(t_{ij})_+^2 = 0$ if $t_{ij} \leq 0$. The random effects, $b_{1i}, b_{2i}, b_{3i}, b_{4i}$, are assumed to have a multivariate normal distribution, with zero mean. In this model, each girl has a separate growth curve that can be described in terms of a linear trend for changes in response before menarche, and a quadratic trend for changes in response after menarche.

The REML estimates of the fixed effects, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)'$, are displayed in Table 13.2. These results suggest that there is significant nonlinearity in the postmenarcheal trend. The estimate of β_4 indicates that increases in percent body fat are greatest around the time of menarche but level off at approximately 4 years following the onset of menarche. The results

Table 13.2 Estimated regression coefficients (fixed effects) and standard errors for the piecewise linear-quadratic model for the percent body fat data

Variable	*Estimate*	*SE*	*Z*
Intercept	20.4201	0.5817	35.10
Time	−0.0155	0.1612	−0.10
$(\text{Time})_+$	4.8439	0.4055	11.94
$(\text{Time})_+^2$	−0.6469	0.0772	−8.38

also suggest that there is no significant increase in percent body fat during the 3 to 4 years prior to menarche.

For this revised model, we consider scatter-plots of the (transformed) residuals versus (transformed) time (see Figure 13.12). The scatter-plots of the transformed and untransformed residuals do not reveal any obvious systematic trends. When lowess smoothed curves are superimposed on the scatter-plots, the curvature that was apparent in Figure 13.8(a) is no longer discernible in Figure 13.12(a). The inclusion of a quadratic trend in the postmenarcheal period has led to an improvement in fit as determined by both the Wald test for the quadratic trend ($Z = -8.38, p < 0.0001$) and the examination of residual diagnostics. Similarly, we can assess the adequacy of the quadratic trend model using cumulative and moving sums of residuals. Figure 13.13 shows a plot of the observed cumulative sum of the residuals, with respect to the covariate time; superimposed on the graph are 10,000 realizations from the Gaussian mean-zero null distribution. This plot suggests there is no systematic trend in the observed curve. This is confirmed by a numerical assessment. The maximum absolute value of the observed cumulative sum is 8.46, with corresponding p-value for the supremum test equal to 0.174. A similar plot can be constructed based on a moving sum and yields the same conclusion. Thus, both the graphical and numerical results suggest that the functional form for time (i.e., piecewise linear-quadratic) is adequate for these data. An overall assessment of linearity (or the link function), assessing the need for a transformation in Y or the mean of Y, can be based on a plot of

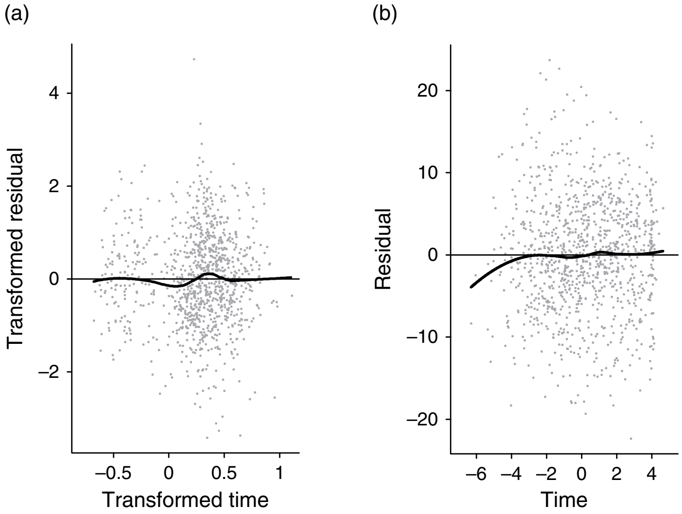


Figure 13.12 Scatter-plot of (a) the transformed residuals versus transformed time, and (b) the untransformed residuals versus time, for the revised model for the percent body fat data

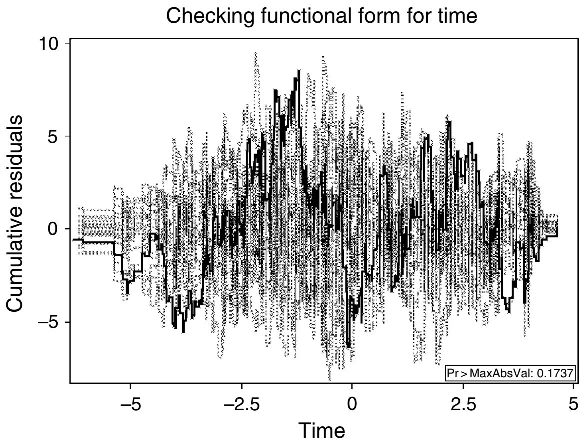


Figure 13.13 Plot of observed cumulative sum of residuals versus time (since menarche) and 10,000 simulated realizations from the null distribution assuming a correctly specified model for mean percent body fat

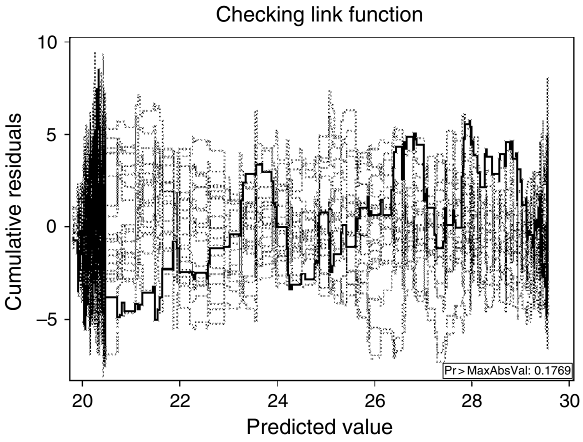


Figure 13.14 Plot of observed cumulative sum of residuals versus the fitted values and 10,000 simulated realizations from the null distribution assuming a correctly specified model for mean percent body fat

the cumulative sum of residuals with respect to the fitted values (see Figure 13.14). This plot also suggests there is no systematic trend in the observed curve; the p-value for the supremum test is equal to 0.177.

Finally, it is worth emphasizing that the graphical and numerical methods based on cumulative and moving sums of residuals are valid regardless of the true joint distribution of the longitudinal response vector; in particular, they do not require correct specification of the covariance among the responses. As such, these graphical and numerical techniques for assessing the model for the mean response are relatively robust to assumptions about the distribution of the responses and assumptions about the covariance among the repeated measures.

4 Conclusion

In this chapter we have reviewed graphical techniques that are useful at both the early and later stages of longitudinal data analysis. We have seen that time plots and smoothed plots of the mean response over time, often stratified by covariates, can be helpful in determining trends in the mean response over time and the appropriate functional form for covariates. Graphical techniques, based on residuals, are especially useful for assessing the adequacy of any postulated model for longitudinal data. They are also useful for identifying observations and individuals that are potential outliers.

Of note, the focus of the graphical and numerical techniques discussed in this chapter has been on the model for the mean response. This reflects the fact that the primary goal of many longitudinal studies is to assess changes in the mean response over time and the factors that influence change. To a large extent, the covariance among repeated measures on the same individuals is regarded as a nuisance characteristic of the data and is of secondary interest. Of course, this does not imply that the covariance can be disregarded or simply ignored. Indeed, the covariance among repeated measures must be properly accounted for to assure valid inferences. Graphical techniques also have a role in the assessment of the adequacy of the variance and correlation assumptions in longitudinal data analysis. For example, the adequacy of the variance assumption can be informally assessed by examining the scatter-plot of the transformed residuals versus the transformed predicted values and/or time. In a correctly specified model for the variance, the range of the transformed residuals should be approximately constant over (transformed) time and for varying $\widehat{\mu}_{ij}^{*}$. A more informative plot is obtained by considering the scatter-plot of the absolute values of the transformed residuals, $|r_{ij}^{*}|$, versus $\widehat{\mu}_{ij}^{*}$ and/or (transformed) time. If the assumed model for the variance is adequate, there should be no systematic trend. An informal check on the overall adequacy of the model for the covariance, both the models for the variances and correlations, is provided by a smoothed plot of the so-called *empirical semi-variogram*. A detailed description of the use of the semi-variogram for longitudinal data can be found in the article by Laird *et al.* (1992), Chapter 10 (Section 10.4) of Verbeke and Molenberghs (2000), Chapter 3 (Section 3.4) of Diggleq *et al.* (2002), and Chapter 9 (Section 9.4) of Fitzmaurice, Laird and Ware (2004).

Software

The transformed residuals discussed in Section 3.2 can be produced as standard output from some statistical packages. For example, they can be obtained using the "normalized" residuals option with the *lme* function in S-PLUS and with the *VCIRY* option with PROC MIXED in SAS. Model checking based on aggregate residuals can be implemented using the *ASSESS* statement in PROC GENMOD in SAS. Because statistical software is constantly evolving, all of the techniques discussed in this chapter should soon be available within most of the major statistical packages.

Acknowledgements

The early sections of this chapter are based in part on Chapter 9 of *Applied Longitudinal Analysis* by Fitzmaurice, Laird and Ware (2004). The writing of this chapter was supported by grants GM 29745, HL 69800, and MH 17119 from the National Institutes of Health.

References

Bandini, L.G., Must, A., Spadano, J.L. and Dietz, W.H. (2002). Relation of body composition, parental overweight, pubertal stage, and race-ethnicity to energy expenditure among premenarcheal girls. *American Journal of Clinical Nutrition*, 76: 1040–1047.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall/CRC Press.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*, 2nd edn. New York: Oxford University Press.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*. New Jersey: Wiley.

Laird, N.M., Donnelly, C. and Ware, J.H. (1992). Longitudinal models with continuous responses. *Statistical Methods in Medical Research*, 1: 225–247.

Lin, D.Y., Wei, L.J. and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, 58: 1–12.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman & Hall/CRC Press.

Naumova, E.N., Must, A. and Laird, N.M. (2001). Evaluating the impact of "critical periods" in longitudinal studies of growth using piecewise mixed effects models. *International Journal of Epidemiology*, 30: 1332–1341.

Phillips, S.M., Bandini, L.G., Compton, D.V., Naumova, E.N. and Must, A. (2003). A longitudinal comparison of body composition by total body water and bioelectrical impedance in adolescent girls. *Journal of Nutrition*, 133: 1419–1425.

Rogan, W.J., Dietrich, K.N., Ware, J.H., Dockery, D.W., Salganik, M., Radcliffe, J., Jones, R.L., Ragan, N.B., Chisolm, J.J. and Rhoads, G.G. (2001). The effect of chelation therapy with succimer on neuropsychological development in children exposed to lead. *New England Journal of Medicine*, 344: 1421–1426.

Treatment of Lead-Exposed Children (TLC) Trial Group (2000). Safety and efficacy of succimer in toddlers with blood leads of 20-44 μg/dL. *Pediatric Research*, 48: 593–599.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Waternaux, C., Laird, N.M. and Ware, J.H. (1989). Methods for analysis of longitudinal data: Blood-lead concentration and cognitive development. *Journal of the American Statistical Association*, 84: 33–41.

Waternaux, C. and Ware, J.H. (1991). Unconditional linear models for analysis of longitudinal data. In Dwyer, J.H., Feinleib, M., Lippert, P. and Hoffmeister, H. (eds), *Statistical Models for Longitudinal Studies of Health*. New York: Oxford University Press.

This page intentionally left blank

Chapter 14

Separating age, period, and cohort effects in developmental and historical research

Scott Menard

This chapter deals with a fundamental issue in longitudinal research, the separation of developmental (age) effects, historical (period) effects, and the effect of experiencing certain historical events at a certain age (cohort effects). Basic to this issue is a discussion of alternative dimensions on which we can measure time in the analysis of change. Section 1 deals with age and period as time dimensions; Section 2 with age, period, and cohort as explanatory variables; and Section 3 with the conceptual status of cohort as a unit of analysis. Section 4 illustrates the dummy variable regression approach to analyzing age, period, and cohort effects, suggesting why its use has declined after some initial popularity. Sections 5 and 6 describe the conceptual approach to analyzing changes over time and age, not only in values of variables but also in relationships among variables, and Section 7 concludes the chapter.

1 Age and period as alternative dimensions of time

In longitudinal research, change is typically measured with reference to one of two continua: chronological time (hereafter simply *time*) or *age*. Time is measured externally to the cases or subjects being studied (e.g., 7:15 p.m., October 26, 2006). Age is measured internally, relative to the subject or case under study (e.g., twenty-five years since birth). The choice of time or age as the underlying continuum for measuring change may be important, and for some purposes it may be useful to consider both in the same analysis. Also important is the distinction between age-related differences when age is measured cross-sectionally (differences between subjects who are 40 years old and subjects who are 50 years old in 1990) and age measured longitudinally (differences between subjects who are 40 years old in 1990 and those same subjects when they are 50 years old in 2000). When age is measured cross-sectionally, the differences between the values of variables for 40-year-olds and the values of variables for 50-year-olds may be interpreted as differences *between* birth cohorts or age groups at a particular time. When age is measured longitudinally, the differences may be interpreted as *developmental* differences *within* a cohort or age group over time.

The demographic definition of a cohort is provided by Glenn (1977, p. 8): “A cohort is defined as those people within a geographically or otherwise delineated population who experienced the same significant life event within a given period of time." A similar definition is offered by Ryder (1965, p. 845): "A cohort may be defined as the aggregate of individuals (within some population definition) who experienced the same event within a given time interval." Both Glenn and Ryder noted that although the term "cohort" is usually used to refer to *birth* cohorts, one may also define cohorts in terms of year of marriage or divorce, year of first employment or retirement, or year of occurrence of other events.

2 Age, period, and cohort as explanatory variables

Hobcraft et al. (1982), in a thorough discussion of age, period, and cohort as explanatory variables, noted that age is "a surrogate—probably a very good one in most applications—for aging or more generally for physiological states, amount of exposure to certain social influences, or exposure to social norms." Although it would be desirable to replace age by the variables for which it is a surrogate, age may generally be expected to perform quite well as an explanatory variable. Indeed, it is possible that age, which can be measured with some precision, may be a more valid measure of such underlying variables than more direct but potentially less reliable measures (e.g., survey measures) of exposure to norms or other social influences. Diagnostic measures of physiological states may be more accurate, but also much more costly, than simply asking a respondent his or her age. Although imperfect, then, age appears to be a reasonable choice as an explanatory variable. Age may be measured cross-sectionally or longitudinally. When it is measured only cross-sectionally, age differences are the same as cohort differences, and the impact of age cannot be separated from the impact of being in a particular cohort. To the extent that we draw conclusions about developmental differences over the life course from purely cross-sectional data, we are assuming that there are no differences associated with being in different cohorts. Mathematically, however, being a certain age and being in a certain cohort are identical, and their effects cannot be separated.

Hobcraft et al. (1982) also assert that " 'Period' is a poor proxy for some set of contemporaneous influences, and 'cohort' is an equally poor proxy for influences in the past. Measured 'effects' of periods and cohorts are thus measures of our ignorance: in particular, of whether the factors about which are ignorant are more or less randomly distributed along chronologically measurable dimensions." If we measure age for a single cohort across multiple periods, being a certain age is mathematically identical to being in a certain period, and the impacts of developmental change (age) and historical change (period) cannot be distinguished. With multiple ages, periods, and cohorts, mathematically cohort (year of birth) = period (calendar year) − age (years since birth), and because the three are linearly dependent, we cannot separate one (linear) effect from the other (in the case in which each is hypothesized to have a linear effect on some outcome). This situation of linear dependence posed a critical problem for the joint analysis of age, period, and cohort effects, because the effect of any one of the variables could, mathematically, just as well be an effect of the other two. For example, an apparently linear decline in fertility or crime could be interpreted as a period effect, or as a combination of age and cohort effects, since period = age + cohort.

In 1973, Mason et al. (1973) developed a dummy variable regression method for parameterizing age, period, and cohort effects, in an effort to overcome the problem of linear dependence among the three variables when age is measured as age at last birthday (integer number of years since birth, e.g., 25), period is measured as current year (e.g., 1990), and cohort, implicitly *birth* cohort, is measured as year of birth (e.g., 1965). Publication of Mason et al.'s (1973) proposed solution to this problem of linear dependence produced three responses. One response was a series of papers that used real and hypothetical data to demonstrate the limitations (particularly sensitivity to assumptions used in model specification) and potentially inappropriate uses of the dummy variable regression technique (Glenn, 1976; 1977; 1981; Greenberg and Larkin, 1985; Rodgers, 1982), and defenses of the robustness and usefulness of the method (Knoke and Hout, 1976; Mason et al., 1976; Smith et al., 1982). A second response was to modify the method somewhat (e.g., Maxim, 1985; Wright and Maxim, 1987) or to develop a method which explicitly attempts to avoid the problems with the dummy variable regression technique by eliminating at least one of the three possible influences (Palmore, 1978). A third response was a proliferation of papers using either the dummy variable regression technique or the technique developed by Palmore to study the effects of age, period, and cohort on a variety of topics, including crime and delinquency (e.g., Lab, 1988; Maxim, 1985; Pullum, 1977; Smith, 1986; Steffensmeier et al., 1987) suicide (e.g., Wasserman, 1987), alcohol and drug use (e.g., O'Malley et al., 1984), fertility (e.g., Wright and Maxim, 1987), divorce (e.g., Carlson, 1979) and other phenomena. After a burst of activity over about a 20-year period, the use of the dummy variable regression approach to the analysis of age, period, and cohort effects has become much less frequent in these areas.

3 Cohort as a unit of analysis

Both Glenn (1997) and Ryder (1965) in their definitions of "cohort" noted that although the term "cohort" is usually used to refer to *birth* cohorts, one may also define cohorts in terms of year of marriage or divorce, year of first employment or retirement, or year of occurrence of any number of other events. Graetz (1987) used the term *event cohorts* to describe cohorts other than birth cohorts. He dealt specifically with cohorts defined in terms of the year of attainment of highest level of education. To the extent that an *event* is not dependent on age or period, an event cohort is not linearly dependent on age or period. As an example, consider the event cohort of year of maximum educational attainment from Graetz (1987). One may terminate one's education with two years of high school at age 16, or with a Ph.D. at age 35. People drop out of school and obtain their doctorates every year, so there is likely to be little relationship between period and the event cohort defined by maximum educational attainment. There is likely to be a nonlinear relationship between age and the end of formal education, with peaks at age 18 and 22 (high school and college graduation), and an increase in the absolute number but not the rate (number/population) of those who end their education in successive years (because of population growth). Note that the event cohort is defined in terms of those who terminate their education at any given level, not, for example, at college graduation, so increasing levels of education may not be clearly reflected in age or period patterns involved in the termination of education.

In the above descriptions and definitions of cohorts, the absence of any linear dependence on age and period for some (not all) event cohorts is a relatively minor point, distinctly secondary in importance to the more fundamental point that *cohorts, as aggregates of individuals, are units of analysis (units upon which measurement is performed) or cases for study*. It is in this sense that the term cohort is primarily used by Ryder (1965), and implemented in some studies (Lloyd et al., 1987; Wetzel et al., 1987; see also Joshi, Chapter 5, and Mayer, Chapter 6, in this volume). Even in some studies which viewed cohort primarily as an explanatory variable, it was also recognized as a unit of analysis (e.g., Wright and Maxim, 1987). An analogy may be drawn between cohorts, which are defined in terms of time and implicitly limited to certain geographic or political boundaries, and aggregations such as census tracts, cities, and nations, which are defined in terms of geographic and political boundaries and implicitly limited to some period in time. An example of the latter would be the 26 American cities for which data on criminal victimization and other variables were collected in the early 1970s (US Department of Justice, 1975a; 1975b; 1976), and upon which several studies of victimization have been based (e.g., Booth et al., 1977; Decker, 1980; Menard and Covey, 1988; O'Brien, 1983). Both cohorts and cities represent aggregates of individuals. Both may be used as cases (rather than variables) in data analysis. Both may have characteristics, such as size and composition, which are aggregate in nature, not reflected in specific individual members of those aggregates. It is these characteristics that are most appropriately treated as variables, rather than cohorts or cities themselves.

Cohorts are aggregates of individuals. Ages and periods are aggregates of years or other units of time. In social and behavioral science research, cohorts have measurable characteristics, some of which are inherently aggregate in nature (size, gender, or ethnic composition) and others which are summations (total number of arrests) or averages (median lifetime income) of the characteristics of the individuals who comprise the cohort. By contrast, we do not measure the aggregate characteristics of ages or periods as such, but instead we measure aggregate characteristics of individuals or other units of analysis *during* a particular period or age. Age 15 and the year 1980 are not differentiated from other ages or years by size (unless you count leap years) or composition; the cohort born in 1965 *is* differentiated from other cohorts by size and composition. Cohorts, which are aggregates of individuals, may be units of analysis. Age and period, which are aggregates of years or other units of time, are variables, and they may be used to delimit the units of analysis for a particular study (e.g., those 15 years old in 1980), but they are not themselves units of analysis in sociological research.

In this respect, cohorts are qualitatively different from ages or periods, neither of which can be described as aggregates of individuals bounded by space and time. Considered in this light, the term "cohort effect" takes a peculiar meaning, and refers to the "effect" of the unit of analysis. By analogy, one could speak of "city effects" when cities are the units of analysis. Operationally, they may be treated in the same way. One may examine differences among cohorts or cities on some set of dependent variables, e.g., in an analysis of variance, or one may use characteristics of the units of analysis such as size or composition to try to explain differences among the units of analysis. The former strategy raises important questions of interpretation, and should almost always lead to the latter strategy. In other words, if we establish that a difference exists (between cohorts or cities), the next step is to explain why that difference exists.

Implementation of the second strategy is not necessarily difficult. According to Ryder (1965), "A cohort's size relative to the sizes of its neighbors is a persistent and compelling feature of its lifetime environment." Mason et al. (1976) noted that age, period, and cohort are proxies for unmeasured variables and indicated that "if cohort size is the variable which causes differentiation in the context of a specific substantive problem, then, if size measurements can be constructed, it is unnecessary to include cohorts as such in the specification because the preferred variable is available." They continued by noting that use of cohort size potentially eliminates the linear dependence problem (unless cohort size increases linearly with time) and hence the problem of estimability for linear regression, and makes the results of the analysis less tentative. Hobcraft et al. (1982) observed that age, period, and cohort (measured as year of birth) are all proxies for other variables, and suggested that for cohort and period in particular, "when these factors can themselves be directly measured, there is no reason to probe for period or cohort effects." Ryder (1965) noted that size is only one of several characteristics which may be used to differentiate cohorts from one another; however, cohort size is the cohort characteristic that has played the most important role in research since the publication in 1968 of Easterlin's work regarding the impact of cohort size on the labor force, and his subsequent work, published in 1980, regarding the impact of cohort size on a variety of social problems, including unemployment, divorce, and crime (Easterlin, 1987). Still, cohort size need not be emphasized to the point that other potentially important cohort characteristics become neglected or excluded. The point is that cohort, measured as year of birth, has sometimes been used when cohort size or some other cohort characteristic would have been more appropriate for studying age, period, and cohort effects. This may stem at least in part from a failure to recognize that age, period, and cohort have qualitatively different statuses as explanatory variables. Although, as noted above, cohort may technically be treated as an explanatory variable, it is generally not appropriate to do so.

For example, some analyses of the Easterlin (1987) relative cohort size hypothesis have first proceeded by calculating a dummy variable regression model (following Mason et al., 1973), then by examining the zero-order correlation between the magnitude of the dummy variable parameters and cohort size (Maxim, 1985; Smith, 1986; Steffensmeier et al., 1987). In this instance, it would have been more appropriate to include cohort size, rather than year of birth, in the original predictive equations. This latter approach, which was used by other researchers (Elliott et al., 1989; Menard, 1992; Menard and Elliott, 1990; Menard and Huizinga, 1989; O'Brien, 1989), has the methodological advantage of eliminating the estimability problem for which the Mason et al. and Palmore techniques were proposed as solutions, and the conceptual advantage of testing the hypothesis directly rather than indirectly. In addition, it may alleviate the aforementioned problem that the dummy variable regression technique may be highly sensitive to assumptions made to identify the model, and may capitalize on chance variation in estimating the model parameters (Greenberg and Larkin, 1975; Rodgers, 1982).

4 Illustration of the dummy variable regression analysis of age, period, and cohort effects

Table 14.1 illustrates these points with a reanalysis from Smith's (1986) study of homicide arrests in which he used the dummy variable regression technique to calculate three models with slightly different assumptions.[1] The zeros in parentheses correspond to dummy variables omitted from each equation in order to identify the model. Also in Table 14.1, ordinary least-squares (OLS) regression results using the same data are presented, but with cohort size in place of year of birth as the variable for cohort effects. From Table 14.1, four major

[1] Smith also used OLS regression equations that included age and cohort size, but not period, as predictors of homicide rates for disaggregated annual age and period specific rates. The present analysis, for purposes of comparison, is limited to the aggregated (five-year interval) data used in the model that included all three types of effects: age, period, and cohort.

Table 14.1 Homicide arrest rates

	Regression coefficients				
	Smith (1986) dummy variable regression			*Continuous variable OLS regression*	
Independent Variable	*Model 1*	*Model 2*	*Model 3*	*Unstandardized*	*Standardized*
Age (all categories)	NA	NA	NA	–.278**	–.546
(15–19)	−.31	−12.86	−6.28		
(20–24)	8.33	−2.15	3.21		
(25–29)	8.33	−.64	4.10		
(30–34)	6.54	.23	3.19		
(35–39)	4.54	.40	2.15		
(40–44)	2.23	(0)	(0)		
(45–49)	(0)	(0)	(0)		
Period (all categories)	NA	NA	NA	.178**	.418
(1952–56)	(0)	(0)	(0)		
(1957–61)	(0)	−1.98	(0)		
(1962–66)	−.07	−4.19	−1.42		
(1967–71)	3.57	−2.62	1.35		
(1972–76)	3.72	−4.54	.62		
Cohort (all: cohort size)	NA	NA	NA	.005*	.380
(1903–07)	(0)	(0)	(0)		
(1908–12)	(0)	(0)	.60		
(1913–17)	−.79	1.99	.89		
(1918–22)	−1.48	3.38	1.08		
(1923–27)	−1.73	5.21	1.72		
(1928–32)	−1.76	7.28	2.72		
(1933–37)	−.25	10.85	4.97		
(1938–42)	1.18	14.36	7.18		
(1943–47)	4.28	19.54	11.27		
(1948–52)	6.88	24.22	14.75		
(1953–57)	8.32	27.74	17.06		
Intercept	5.76	7.16	6.10	−3.88	NA
Explained variance (R^2)	.99	.99	.99	.47	.47

* Significant at the .05 level
** Significant at the .01 level

points should be made. First, with regard to the existence of nonlinear effects, the three models give inconsistent results for age and period. The peak age for homicide arrests (allowing for period and cohort effects) is 20–29 for Model 1, 35–39 for Model 2, and 25–29 for Model 3. Such instability might be attributed to the small number of ages and years, but Rodgers (1982) indicated that the dummy variable regression technique is typically sensitive to the assumptions made to identify the model, and such instability across models with different assumptions is to be expected. For cohort, Model 1 shows a decline followed by an increase in the cohort parameters, but Models 2 and 3 show monotonic increases in the cohort parameters. The cohort parameters are highly correlated with cohort size: .88 for Model 1, .74 for Model 2, and .81 for Model 3. Cohort size thus explains 55–78% of the variance in the effects of cohorts, measured as year of birth. Briefly, then, the first point is that the dummy variable regression technique does not necessarily provide a reliable guide to identifying or describing the form of the relationship (linear as opposed to nonlinear effects). The dummy variable regression results are not unique for a given set of variables. Instead, as Smith (1986) demonstrated, the results vary, depending on which categories are excluded or set equal to one another in order to identify the model.

The second major point is that the substantive conclusions from the dummy variable regression and OLS regression approaches may differ. Using the dummy variable approach, Smith concluded that cohort effects appear to be strongest. In the OLS approach (using cohort size instead of year of birth), the age effect appears to be strongest and the cohort size effect appears to be weakest, based on the standardized regression coefficients. All of the effects in the OLS equations are statistically significant at the .05 level, and the age and period effects are statistically significant at the .01 level.[2] Use of the dummy variable regression technique, then, will not necessarily produce the same results as use of continuous variables in a regression analysis.

The third major point is that the results of the OLS approach with continuous variables are more readily interpretable than those obtained with the dummy variable regression approach. In the dummy variable regression analysis, the units of measurement are defined in terms of the omitted categories. In the OLS regression with continuous variables, the units of measurement are defined, not in terms of omitted categories, but in terms of more natural units of measurement: years (actually, five-year intervals in Smith's data) for age and period, arrests per 100,000 population for the dependent variable, and thousands of births for cohort size. Based on the results using cohort size, we can say that the homicide arrest rate declines .278 arrests per 100,000 people (or about 3 per million) for every five-year increase in age; that it increases at a rate of .178 arrests per 100,000 people (about 2 per million) every five years (historically); and that it increases 5 arrests per 100 million people for every increase of 1000 births in a cohort.

A fourth point is that the results using cohort size make more intuitive sense than those obtained using year of birth. First, the number

[2] The use of statistical significance tests here is consistent with the recommendation of Winch and Campbell (1969), who suggested the use of significance tests even with population data to allow us to evaluate the possibility that apparent relationships in the data are the result of haphazard variation in the data (chance, instability) rather than systematic relationships among variables. For a dissenting view of this use of significance tests, see Morrison and Henkel (1969). This debate lies outside the focus of the present chapter, and the results regarding statistical significance may be ignored without changing the conclusions reached here.

of parameters is more manageable: three in the equation using cohort size, and 23 in the equation using year of birth. Second, the explained variance appears to be inflated in the dummy variable regression equations. This may be attributable at least in part to the fact that the number of parameters (23) is large relative to the number of values of the dependent variable (35). The situation becomes worse for larger numbers of ages and periods; as the number of ages and periods increases, the number of parameters to be estimated approaches the number of values of the dependent variable. Ideally, in regression analysis, one should have several times as many values of the dependent variable as there are parameters to be estimated. Otherwise, the parameter estimates capitalize on chance variation, and the explained variance is inflated (Johnston, 1984; Kleinbaum, Kupper and Muller, 1988). The 99% explained variance in the dummy variable regression model is not so much an indicator of how well historical, developmental, and cohort effects have an impact on aggregate homicide arrest rates as it is a reflection of the purely statistical properties of the dummy variable regression "accounting system" for age, period, and cohort effects; a comparable level of explained variance can be expected in such dummy variable regression models regardless of the substantive impact of the variables in question. The 47% explained variance in the OLS equation with cohort size, in contrast, is more plausible as an estimate of the combined impact of developmental, historical, and cohort size influences on the homicide arrest rate.

The point of all this is that cohort, measured as year of birth, has sometimes been used when cohort size or some other cohort characteristic (or a nonlinear interaction term involving age and period) would have been more conceptually or theoretically appropriate for studying age, period, and cohort effects. This may stem at least in part from a failure to recognize that age, period, and cohort have qualitatively different conceptual statuses. Although, as noted above, cohort membership may be treated as an explanatory variable from a purely methodological viewpoint, theoretically and substantively it is not generally appropriate to do so. Age and period are more appropriate as explanatory variables, age more so than period (Hobcraft et al., 1982). Ideally, one would eliminate period and cohort and replace them with the variables for which they act as proxies in any causal analysis.

5 Period effects: Changes over time

If we are concerned only with changes over chronological time (historical changes) and not with changes over age (developmental changes), we must either be certain that age is entirely irrelevant, or include age as an explanatory variable, or control for age by making age-specific comparisons. One of the safest ways to approach the study of trends over time is to use age-specific comparisons. In an age-specific comparison, only those cases of a certain age in one year are compared with cases of the same age in some subsequent year. The age may represent a single year (e.g., age 15) or a range of years (e.g., over age 65), and separate comparisons may be made for all possible ages or age groups. Some variables are naturally age-specific, e.g., Scholastic Aptitude Test (SAT) scores (because the SAT is taken primarily by individuals aged 16–18) or in infant mortality rates. In other instances, it is necessary to explicitly control for age. For example, Gold and his associates (Gold and Reimer, 1975; Williams and Gold, 1972) examined rates of self-reported delinquency among national probability samples of 13–16-year-olds in a repeated cross-sectional design, and found little evidence of change from 1967 to 1972. Menard (1987) obtained similar results for national probability samples of 15–17-year-olds from 1976 to 1980. Covey and Menard (1987; 1988) examined trends in victimization and trends in arrests for those over age 65, and found that rates of arrests were generally increasing and rates of victimization were generally decreasing among this older age group.

In each of the above-cited studies, controlling for age produced relatively unambiguous evidence for the existence or absence of trends over time. Without such controls for age, it may be difficult to ascertain whether changes are historical or developmental in nature, even if the entire population is used instead of a sample. Chilton and Spielberger (1971) examined changes in official crime rates, and found that much of the apparent change over time (what appeared on the surface to be a change in behavior) was attributable to changes in the age structure, or, more specifically, to changes in the percentage of the population in the adolescent ages. Individual studies will vary, but in general it is appropriate to consider the possibility that apparent period trends may actually be attributable to changes in age (at the individual level) or age composition (at the aggregate level).

Another concern of longitudinal research is the examination of changes, not in values or levels of variables over time, but in *relationships* between or among variables over time. It is one thing, for example, to say that mortality has been declining for over two centuries. It is another to indicate that in the early stages of mortality decline, reductions in mortality were achieved primarily by public health measures (sanitation, access to safe drinking water, pasteurization, etc.) and medicine played little if any role, but in the later stages of the decline, advances in medicine (inoculation, antibiotics) rather than public health measures were responsible for mortality declines (McKeown, 1976; McKeown and Record, 1962; McNeill, 1976). Hout et al. (1999) examined the relationship between social class (six categories, from professional to less skilled blue collar) and voting behavior in American presidential elections from 1944 to 1992, and found different patterns for different classes. The tendency of the highest (professional) class to support Republican candidates increased over time, while the tendency of the three lower socioeconomic classes to vote Democratic declined over time, particularly for the nonprofessional self-employed and the skilled manual classes.

6 Age effects: life cycle and developmental changes

Baltes and Nesselroade (1979) listed five objectives or rationales for longitudinal (or more specifically, in their case, prospective panel) research: (1) direct identification of intraindividual change, i.e., whether individuals change from one period to another; (2) direct identification of interindividual similarities or differences in intraindividual change, i.e., whether individuals change in the same or different ways; (3) analysis of interrelationships in behavioral change, i.e., whether certain changes are correlated with each other; (4) analysis of causes or determinants of intraindividual change, i.e., why individuals change from one period to another; and (5) analysis of causes or determinants of interindividual similarities or differences in intraindividual change, i.e., why different individuals change in different ways from one period to another. All of these objectives are concerned with patterns of developmental change, specifically at the individual level, although they are easily extended to aggregate levels (groups, organizations, cities, nations). At the individual level, intraindividual changes may include things people think (becoming more politically conservative), things they do (becoming employed, changing jobs, retiring), or things that are done to them (being arrested or being robbed). In the study of intraindividual change, age serves as a proxy for age-related physiological changes and exposure to social influences (Hobcraft et al., 1982) which may be difficult or costly to measure directly.

For some purposes it may be reasonable to draw simple inferences about intraindividual change from cross-sectional data. For example, from cross-sectional data on rates of arrest and childbearing by age, we may reasonably infer that one's likelihood of being arrested or of having a baby is practically nonexistent before age 7, increases in adolescence and early adulthood, and diminishes substantially after age 65. There would seem to be little chance that these age-related differences may be explained by period effects or cohort characteristics. On the other hand, it would not be safe to infer that people become more conservative and less educated as they get older, based on cross-sectional data. Age differences in political attitudes at a particular period may reflect either changes in attitudes with age, or constancy in attitudes over the life cycle coupled with differences in attitudes between cohorts. If older people have less education than younger people, it is not because they become "de-educated"; a more plausible explanation is that educational attainment has increased over time (a period effect), resulting in differences in the average educational level of successive cohorts.

A more compelling need for longitudinal data arises if we wish to study "career" patterns of behavior. The most obvious application of this is in the study of labor market careers, from initial job entry through patterns of promotion, job change, job loss, and eventually either retirement or death. Closely related to this is the study of status attainment careers, which includes consideration of educational attainment as well as occupational status and income (e.g., Blau and Duncan, 1966). Other applications of the "career" perspective include marital histories (e.g., Becker et al., 1977), educational attainment and the process of learning (e.g., Heyns, 1978), and criminal careers (e.g., Blumstein et al., 1986). Such studies have in common a concern with patterns of entry, continuity, and exit from the behavior upon which the career is based, and with the correlates and potential causes associated with changes or discontinuities in the behavior (unemployment and obtaining a new job; divorce and remarriage; dropout and re-entry in education; suspension and resumption of criminal behavior). It is only with longitudinal data, and more specifically panel data, that many of the questions regarding developmental career patterns may be answered.

Life course research (Giele and Elder, 1998) is similar to the study of individual careers, but broadens the career paradigm to explicitly locate intraindividual change within a broader historical and social context. Integral to the life course perspective are issues of (1) location in time (history) and place (society and culture); (2) linked lives, the integration of individuals' lives with one another at the interpersonal and social institutional levels; (3) human agency, the ability and tendency of individuals to set goals and decide how to pursue them; and (4) timing of lives, individuals' making decisions about whether and when to act in certain ways, or formulating strategies for living based not only on internalized goals, but also in response to external events or conditions. In contrast to perspectives that see life transitions as progressing through a fixed sequence of stages, the life course perspective recognizes the interindividual variation in the sequencing of life transitions as responses to differences in individual goals (human agency) and external influences (timing of lives). Life course research focuses on phenomena which can be adequately analyzed only with long-term longitudinal research: event histories or trajectories which differ across individuals in timing, duration, or rates of change.

Parallel to the earlier concern with examining changes in the strength or pattern of relationships from one period to another, we may want to examine changes in the strength or pattern of relationships from one age to another. Here again the issue of whether to base the comparison on cross-sectional (intercohort) or longitudinal (intracohort) data may arise, and as before the decision hinges on whether we are concerned with how well the developmental changes are reflected in the cross-sectional data. If longitudinal data are used, the issue of whether any change that occurs is attributable to age, period, or cohort effects must again be considered.

Using data from the National Youth Survey, a prospective longitudinal panel survey of respondents aged 11–17 in 1976 and 21–27 in 1986, Menard et al. (1989) found that marriage during adolescence was positively associated with substance use and mental health problems, but that marriage during young adulthood (ages 21–27) was negatively associated with substance use and mental health problems. Being enrolled in school had a negative association with illegal behavior, substance use, and mental health problems in adolescence, but no association with illegal behavior, substance use, or mental health problems for young adults. Wofford (1989), analyzing the same sample, found that employment was associated with higher rates of serious illegal behavior in adolescence and lower rates of serious illegal behavior in young adulthood (ages 18–24 in this study). Substantively, these results require explanation. From a life course perspective, there may be age-specific norms for certain behaviors (school, marriage, work), and violating those norms may place one at greater risk of involvement in illegal or problem behavior. Methodologically, these results suggest that relationships among variables may change over the life course, and that it may be appropriate to test for the existence of such changes. In addition, the explanatory power of different theories may vary across the life course. In a series of tests of strain and control theories of illegal behavior, Menard (1992; 1997; Menard et al., 1993) found that the explanatory power of the theories was uniformly weakest in early adolescence (ages 11–14) and stronger in later adolescence (ages 14–17) and early adulthood (ages 17–20) across a range of behaviors from minor delinquency and marijuana use to serious delinquency and hard drug use. With cross-sectional data, such differences may be attributed to age or to intercohort differences; with longitudinal data on multiple cohorts, it becomes possible to estimate the extent to which the differences are developmental, as opposed to period or intercohort differences.

7 Conclusion

Selection of the time dimension is a necessary first step in longitudinal research, and there may be more than one time dimension of interest, particularly in the study of individual as opposed to aggregate change over time (although it is also possible, as illustrated above, that more than one time dimension will be of interest for aggregate data as well). To the extent that longitudinal research involves multiple age groups over multiple time periods, separation of developmental and historical effects may be of interest, and for longer time periods and age spans, consideration of the impact of the interaction of age and period effects (being a certain age at a certain time) may complicate the analysis of developmental and historical trends. One approach is to treat cohorts as units of analysis, analyzing them separately and noting the qualitative and quantitative differences among them in the explanation of differences in developmental and historical effects across cohorts. Another approach is to attempt to identify those characteristics of cohorts that may influence either historical and developmental outcomes, e.g., cohort size-effects on unemployment rates), or may influence the impacts of history and development on those outcomes (e.g., cohort size effects on the relationship between unemployment rates and criminal behavior), and incorporate them in the analysis.

Based on the foregoing discussion, we may conclude the following: (1) Age is an appropriate explanatory variable, but not a unit of analysis. In some instances it may be appropriate to replace age with other variables for which age is a proxy, but often using age itself is the most reliable and cost-efficient option. (2) Period is weakly appropriate as an explanatory variable, but not a unit of analysis. It will often be appropriate to replace period with either the specific historical event of interest (e.g., the Iraqi invasion of Kuwait), or better still with an indicator of *exposure* to that specific event (e.g., whether one was directly involved in the invasion or its aftermath, related to or knew someone directly involved, or merely aware of the event). (3) Cohort is appropriate as a unit of analysis, but is a poor or inappropriate explanatory variable. Cohort as year of birth is best replaced by the relevant characteristic of the cohort (e.g., cohort size), or the relevant *age-specific* experiencing of or exposure to a specific event (e.g., being 18 or 81 during the Iraqi invasion of Kuwait, or being age 18 or 81 for a more direct measure of exposure to the Iraqi invasion of Kuwait, as described above). Use of these guidelines allows us to better examine developmental and historical influences on behavior, and to sort out developmental from historical from cohort-based influences on behavior.

Author's note

Parts of this chapter were taken and/or adapted from my previous work, particularly Menard (2002); other parts are original to this chapter.

References

Baltes, P. B. and Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade and P. B. Baltes (eds), *Longitudinal Research in the Study of Behavior and Development*, pp. 1–39. New York: Academic Press.

Becker, G. S., Landes, E. M. and Michael, F. T. (1977). An economic analysis of marital instability. *Journal of Political Economy*, 82: 1141–1187.

Blau, P. M. and Duncan, O. D. (1966). *The American Occupational Structure*. New York: Wiley.

Blumstein, A., Cohen, J., Roth, J. A. and Visher, C. A. (eds) (1986). *Criminal Careers and "Career Criminals"*, Volumes 1 and 2. Washington, DC: National Academy Press.

Booth, A., Johnson, D. R. and Choldin, H. M. (1977). Correlates of city crime rates: Victimization surveys versus official statistics. *Social Problems*, 25: 187–197.

Carlson, E. (1979). Divorce rate fluctuation as a cohort phenomenon. *Population Studies*, 33: 523–536.

Chilton, R. and Spielberger, A. (1971). Is delinquency increasing? Age structure and the crime rate. *Social Forces*, 42: 487–493.

Covey, H. C. and Menard, S. (1987). Trends in arrests among the elderly. *Gerontologist*, 22: 666–672.

Covey, H. C. and Menard, S. (1988). Trends in elderly criminal victimization from 1973–1984. *Research on Aging*, 12: 329–341.

Decker, S. H. (1980). *Criminalization, Victimization, and Structural Correlates of Twenty-Six American Cities*. Saratoga, CA: Century Twenty-One.

Easterlin, R. A. (1987). *Birth and Fortune*, 2nd edn. Chicago: University of Chicago Press.

Elliott, D. S., Huizinga, D. and Menard, S. (1989). *Multiple Problem Youth*. New York: Springer-Verlag.

Giele, J. Z. and Elder, G. H., Jr. (eds) (1998). *Methods of Life Course Research: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.

Glenn, N. (1976). Cohort analysts' futile quest: Statistical attempts to separate age, period, and cohort effects. *American Sociological Review*, 41: 900–904.

Glenn, N. (1977). *Cohort Analysis*. Beverly Hills, CA: Sage.

Glenn, N. (1981). Age, birth cohorts, and drinking: An illustration of the hazards of inferring effects from cohort data. *Journal of Gerontology*, 36: 362–369.

Gold, M. and Reimer, D. J. (1975). Changing patterns of delinquent behavior among Americans 13 through 16 years old: 1967–1972. *Crime and Delinquency Literature*, 2: 483–577.

Graetz, B. (1987). Cohort changes in educational inequality. *Social Science Research*, 16: 329–344.

Greenberg, D. F. and Larkin, N. J. (1985). Age-cohort analysis of arrest rates. *Journal of Quantitative Criminology*, 1: 227–240.

Heyns, B. (1978). *Summer Learning and the Effects of Schooling*. New York: Academic Press.

Hobcraft, J., Menken, J. and Preston, S. (1982). Age, period, and cohort effects in demography: A review. *Population Index*, 48: 4–43.

Hout, M., Manza, J. and Brooks, C. (1999). Classes, unions, and the realignment of US presidential voting, 1952–1992. In G. Evans (ed.), *The End of Class Politics: Class Voting in Comparative Perspective*, pp. 83–95. Oxford, UK: Oxford University Press.

Johnston, J. (1984). *Econometric Methods*, 3rd edn. New York: McGraw-Hill.

Kleinbaum, D. G., Kupper, L. L. and Muller, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods*, 2nd edn. Boston: PWS-Kent.

Knoke, D. and Hout, M. (1976). Reply to Glenn. *American Sociological Review*, 41: 905–908.

Lab, S. P. (1988). Analyzing change in crime and delinquency rates: The case for cohort analysis. *Criminal Justice Research Bulletin*, 3: 1–8.

Lloyd, L. K., Armour, P. and Smith, R. J. (1987). Suicide in Texas: A cohort analysis of trends in suicide rates, 1945–1980. *Suicide and Life-Threatening Behavior*, 17: 205–217.

Mason, W. M., Mason, K. O. and Winsborough, H. H. (1976). Reply to Glenn. *American Sociological Review*, 41: 904–905.

Mason, K. O., Mason, W. M., Winsborough, H. H. and Poole, W. K. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review*, 38: 242–258.

Maxim, P. S. (1985). Cohort size and juvenile delinquency: A test of the Easterlin hypothesis. *Social Forces*, 63: 661–681.

McKeown, T. (1976). *The Modern Rise of Population.* London: Edward Arnold.

McKeown, T. and Record, R. (1962). Reasons for the decline of mortality in England and Wales during the 19th century. *Population Studies*, 12: 94–122.

McNeill, W. H. (1976). *Plagues and Peoples.* New York: Anchor/Doubleday.

Menard, S. (1987). Short-term trends in crime and delinquency: A comparison of UCR, NCS, and self-report data. *Justice Quarterly*, 2: 11–55.

Menard, S. (1992). Demographic and theoretical variables in the age-period-cohort analysis of illegal behavior. *Journal of Research in Crime and Delinquency*, 29: 178–199.

Menard, S. (1995). A developmental test of Mertonian anomie theory. *Journal of Research in Crime and Delinquency*, 32: 136–174.

Menard, S. (1997). A developmental test of Cloward's differential-opportunity theory. In N. Passas and R. Agnew (eds), *The Future of Anomie Theory*, pp. 142–186. Boston: Northeastern University Press.

Menard, S. (2002). *Longitudinal Research.* Thousand Oaks, CA: Sage.

Menard, S. and Covey, H. C. (1988). UCR and NCS: Comparisons over space and time. *Journal of Criminal Justice*, 16: 371–384.

Menard, S. and Elliott, D. S. (1990). Self-reported offending, maturational reform, and the Easterlin hypothesis. *Journal of Quantitative Criminology*, 6: 237–267.

Menard, S., Elliott, D.S. and Huizinga, D. (1989). The dynamics of deviant behavior: A national survey progress report. National Youth Survey Report No. 49. Boulder, CO: Institute of Behavioral Science.

Menard, S., Elliott, D. S. and Wofford, S. (1993). Social control theories in developmental perspective. *Studies on Crime and Crime Prevention*, 2: 69–87.

Menard, S. and Huizinga, D. (1989). Age, period, and cohort size effects on self-reported alcohol, marijuana, and polydrug use: Results from the National Youth Survey. *Social Science Research*, 18: 174–194.

Morrison, D. E. and Henkel, R. E. (1969). Significance tests reconsidered. *American Sociologist*, 4: 131–140.

O'Brien, R. M. (1983). Metropolitan structure and violent crime: Which measures of crime? *American Sociological Review*, 48: 434–437.

O'Brien, R. M. (1989). Relative cohort size and age-specific crime rates: An age-period-relative-cohort-size model. *Criminology*, 27: 57–78.

O'Malley, P. M., Bachman, J. G. and Johnston, L. D. (1984). Period, age, and cohort effects on substance use among American youth, 1976–1982. *American Journal of Public Health*, 74: 682–688.

Palmore, E. (1978). When can age, period, and cohort be separated? *Social Forces*, 57: 282–295.

Pullum, T. W. (1977). Parameterizing age, period, and cohort effects: An application to US delinquency rates, 1964–1973. In K. F. Schuessler (ed.), *Sociological Methodology 1978*, pp. 116–140. San Francisco: Jossey-Bass.

Rodgers, W. L. (1982). Estimable functions of age, period, and cohort effects. *American Sociological Review*, 47: 774–787.

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review*, 30: 843–861.

Smith, M. D. (1986). The era of increased violence in the United States: Age, period, or cohort effect?, *Sociological Quarterly*, 27: 239–251.

Smith, H. L., Mason, W. M. and Fienberg, S.E. (1982). More chimeras of the age-period-cohort accounting framework: Comment on Rodgers. *American Sociological Review*, 47: 787–793.

Steffensmeier, D., Streifel, C. and Harer, M. D. (1987). Relative cohort size and youth crime in the United States, 1953–1984. *American Sociological Review*, 52: 702–710.

US Department of Justice. (1975a). *Criminal Victimization Surveys in 13 American Cities*. Washington, DC: US Government Printing Office.

US Department of Justice. (1975b). *Criminal Victimization Surveys in the Nation's Five Largest Cities*. Washington, DC: US Government Printing Office.

US Department of Justice. (1976). *Criminal Victimization Surveys in Eight American Cities: A Comparison of 1971/72 and 1974/75 Findings*. Washington, DC: US Government Printing Office.

Wasserman, I. M. (1987). Cohort, age, and period effects in the analysis of US suicide patterns: 1933–1978. *Suicide and Life-Threatening Behavior*, 17: 179–193.

Wetzel, R. D., Reich, T., Murphy, G. E., Province, M. and Miller, J. P. (1987). The changing relationship between age and suicide rates: Cohort effect, period effect, or both? *Psychiatric Developments*, 3: 179–218.

Williams, J. and Gold, M. (1972). From delinquent behavior to official delinquency. *Social Problems*, 22: 209–229.

Winch, R. F. and Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *American Sociologist*, 4: 140–143.

Wofford, S. (1989). A preliminary analysis of the relationship between employment and delinquency/crime for adolescents and young adults. National Youth Survey Report No. 50. Boulder, CO: Institute of Behavioral Science.

Wright, R. E. and Maxim, P. S. (1987). Canadian fertility trends: A further test of the Easterlin hypothesis. *Canadian Review of Sociology and Anthropology*, 24: 339–357.

Chapter 15

An introduction to pooling cross-sectional and time series data

John L. Worrall

1 Introduction

Pooling time series and cross-sectional data simply amounts to gathering repeated observations on several units of analysis. Unfortunately, "pooling" is not the only term used to describe such research designs. Pooled datasets are most frequently called "panel datasets." Other researchers prefer to describe their data as being of the "time series-cross section" (TSCS) variety. Still others use the term "multiple time series." There is some disagreement in the literature as to which term should apply, and when. For example, some have argued that panel data consist of relatively few observations on several units of analysis. In contrast, TSCS data contain many observations on relatively few units (e.g., Beck and Katz, 1995). Others, though, have argued that the same issues and methodological concerns present themselves regardless of how many units and/or time periods are included in a dataset (e.g., Kristensen and Wawro, 2003). The latter view is taken here.

2 Three pooling problems

At first glance, pooling time series and cross-section data would seem advantageous. A researcher who had data on, say, 50 units could do little in terms of quantitative analysis. By adding a hypothetical 10 time periods to each of those units, however, the sample size suddenly increases tenfold. Indeed, this increase in sample size is one of the key advantages of pooling. But pooling raises a number of issues that frequently-used statistical techniques, such as ordinary least squares (OLS) regression, are unequipped to address. The most significant of those issues is the addition of a time dimension.

OLS applied to a cross-sectional dataset requires no concerns with autocorrelation, a problem that occurs when data are not independent along the time dimensions. By definition, cross-sectional data have no time dimension. However, when repeated observations are gathered for the same units, researchers cannot ignore serial dependence in the data. We know, for example, that an individual's behavior is closely tied to his or her previous behavior. Moving to the macro level, we know that certain government activities (such as budget allocations) are closely correlated from one time period to the next.

Also important is a variation of a problem that routinely appears in the OLS context: Heteroskedasticity. A key OLS assumption is homoskedastic errors. When error variance is not constant across units, the result is heteroskedasticity, and steps must be taken to correct for it. When time series and cross-section data are pooled, a potential result is *panel* heteroskedasticity. This occurs when the error variance varies across units (over time) due to characteristics unique to each unit. In the "ordinary" heteroskedasticity context, some units are more variable than others. The effects, though, may be relatively modest, as such heteroskedasticity affects only one unit at a time. But when a time dimension is added, the effects can magnify in a manner equivalent to the number of time periods (Stimson, 1985, p. 919).

The third problem that arises in pooling time series and cross sections is known as heterogeneity. This can occur when all units are affected by a "shock" during the same time period. A re-election, a downturn in the economy, or some other sudden event could cause such a problem. More formally, the errors across each unit will be correlated due to the event they all experienced. Another form of heterogeneity consists of time-stable differences between units. Once these sorts of problems are addressed, the resulting estimates can be interpreted just like ordinary least squares estimates.

3 Fixed effects or random effects: three considerations

Two basic methods are available for analyzing panel data. For the sake of full disclosure, there are *scores* of techniques available (and continually being developed), but most are built around some variation of the two basic choices: Fixed effects models and random effects models. We begin with the basic cross-sectional OLS model, which looks something like this:

$$y = \alpha + \beta x + \varepsilon \qquad (1)$$

The notation for (1) is in matrix form. The "y" is the dependent variable, "α" is the intercept, "x" is a vector of independent variables, "β" represents the regression coefficients, and "ε" is the error term. The pooled model simply extends (1) to:

$$y_{i,t} = \alpha + \beta x_{i,t} + \varepsilon_{i,t} \qquad (2)$$

In (2) the "i" and "t" subscripts denote that we have pooled observations over units "i" and time periods "t." Note that there is nothing different between (1) and (2) other than the fact that (2) acknowledges there are repeated observations on the same units. (2) is sometimes called the "constant coefficients model," implying that the regression coefficients are constant across units and time periods.

The problem with (2) is that even though it acknowledges repeated observations on the same units, it ignores that fact. In other words, (2) makes no attempt to model the repeated observations. The unit and time dimensions therefore need to be taken into account. This is where the fixed and random effects models become viable options for dealing with the panel data structure. The fixed effects models extend (2) to:

$$y_{i,t} = \alpha_i + \delta_t + \beta x_{i,t} + \varepsilon_{i,t} \qquad (3)$$

Note that (3) is the same as (2), but that (3) adds separate intercepts for each unit (denoted by "α_i"). Also note the addition of "δ_t." The "t" subscript denotes dummy variables for each time period. The logic for including separate intercepts for each unit and dummy variables for each time period will be discussed later because our concern here is with the differences between fixed and random effects estimation.

The random effects model looks like this:

$$y_{i,t} = \alpha + \beta x_{i,t} + u_i + w_t + \varepsilon_{i,t} \qquad (4)$$

Note two important differences between (3) and (4). (4) adds u_i and w_t, both of which specify separate errors terms for both unit and time period. Also note the removal of the subscript "i" from α. This model assumes that unobserved differences between units and time are random variables, compared with the assumption included in equation (3) that they are fixed. (4) is commonly called an error components model.

What, then, would lead a researcher to choose between fixed or random effects estimation? There are at least three considerations: (1) the size of N and T; (2) correlation between the error term and observed variables; and (3) variation over time in the predictors. Each of these considerations is touched on in the next three subsections. The fourth subsection briefly introduces a more objective approach to deciding between fixed and random effects estimation, the use of a Hausman test.

3.1 Size of *N* and *T*

With respect to fixed effects models, when the number of observations and/or time periods becomes large and unwieldy, degrees of freedom are sacrificed and efficiency is lost. Fixed effects regressions of data with repeated observations on, say, 5000 cities would require the addition of 4999 additional parameters to the model—simply to model city-specific effects. Add to that a number of time periods and efficiency suffers even more. If N and T are large, that does not automatically mean a researcher should abandon fixed effects estimation. Rather, some additional considerations, each related to efficiency, ought to weigh in to the decision.

One such consideration is complicated interpretation. The researcher may wish, for substantive reasons, to examine the parameter estimates for each unit (or time period). If such information is of no substantive concern, then perhaps a random effects approach would be preferable. Likewise, multicollinearity should be considered. The estimation of multiple parameters is often confounded by correlation among predictor variables. Random effects estimation may be preferable if multicollinearity is a problem.

Fixed effects estimation also sacrifices some possibly useful information. In particular, it removes any of the average unit-to-unit variation from the analysis. The introduction of fixed effects for each unit, for example, simply asks whether intraunit changes in some dependent variable are associated with intraunit changes in one or more independent variables. In other words, fixed effects estimation ignores the possibility that unit-to-unit variation sheds light on the relationship between x and y.

3.2 No correlation between the error term and predictors

It would seem that efficiency losses associated with fixed effects estimation would drive one to opt for random effects estimation. But random effects estimation suffers from certain faults and limitations, as well. In particular, it assumes that the error term is not associated with any of the predictor variables. In other words, the assumption is that the predictor variables are not correlated with unobserved unit-specific effects. Most researchers are hard-pressed to make such a case.

Random effects models further assume that the random error terms are unique to each unit of analysis (see (4) above) and do not change over time. Why would something predictive of the outcome materialize at one point in time and then remain constant (see Berk, 2004, pp. 178–180)? Random effects estimation can also be desirable, however, when one key limitation of the fixed effects approach manifests itself. To that we now turn.

3.3 Problematic predictors

With respect to fixed effects estimation, there are three types of “problematic” predictor variables that limit its use. The first is a predictor variable that does not vary over time, such as a variable denoting whether a person, city, or county is liberal or conservative. Such a variable would be perfectly collinear with dummies for each unit. Likewise, predictor variables that model events every unit experiences at the same time are perfectly correlated with the time dummies (see Allison, 1994). This does not mean, however, that events cannot be modeled in the fixed effects context. A later section of this chapter discusses modeling events with panel data.

Predictors that change little over time are also problematic. Subtle changes from one year to the next can make a predictor variable look like a constant. As Beck and Katz (2004) have observed, "Fixed effects clearly eliminates any stable variables from the analysis, but also makes it difficult for variables that change only slowly to show their impact (when their impact is by and large inter- and not intraunit)." Researchers must then opt either for random effects estimation or, as we will see below, more sophisticated measures of interventions that are otherwise operationalized as dichotomous variables.

3.4 The Hausman test

Sometimes it is easier to choose between fixed and random effects estimation by using a Hausman test (see Hausman, 1978). This is a test of the null hypothesis that random effects would be consistent and efficient against the alternative hypothesis that random effects would be inconsistent. The question is whether there is significant correlation between the unobserved unit-specific random effects and the regressors. If there is no correlation, then the random effects model may be more powerful and parsimonious.

The test statistic is calculated as $[(\text{ß}_{FE}\text{-} \text{ß}_{RE})/(s^2{}_{\beta\text{ß}FE}\text{-}\ s^2{}_{\text{ß}RE})]$ where ß_{FE} are the fixed effects model coefficients, ß_{RE} are the random effects model coefficients, $s^2{}_{\text{ß}FE}$ and $s^2{}_{\text{ß}RE}$ are the variances of the fixed and random effects model coefficients, and the statistic has a chi-square distribution with as many degrees of freedom as there are predictors in the model. An insignificant p-value (greater than .05) means it is safe to use random effects. If the p-value is significant, however, fixed effects should be used. While it is tempting to take the test statistic at face value, as an "objective" criterion for choosing between one or the other approach, common sense should also be used. Substantive reasons may instead drive a researcher to choose random in lieu of fixed effects—or vice versa. Researchers should weigh all of the considerations raised throughout this section.

4 Estimation issues in fixed effects models

The focus of the rest of this chapter will be on fixed effects models, in contrast to random effects models. Fixed effects models are more commonly used in aggregate-level models where researchers routinely estimate the effects of events such as new policies and legal interventions. Recall, though, that there are situations where fixed effects regressions should not be used (such as with constant predictor variables).

There are five key estimation issues associated with fixed effects regression models. Save for the first, the others also apply in the random effects context. The issues are (1) heterogeneity; (2) dynamics; (3) panel heteroskedasticity and contemporaneous correlation; (4) stationarity; and (5) trends. Pooling of time series and cross-section data often causes various combinations of these issues/problems to arise.

4.1 Heterogeneity

Heterogeneity refers, generally, to differences in the units of analysis. For example, in a study of California counties it is clear that there are differences from one county to the next that need to be modeled. Differences of this sort are no less apparent in the cross-sectional context, but researchers cannot expressly model such heterogeneity given the lack of a time dimension. When a time dimension is added, however, researchers can expressly model unit-specific heterogeneity by estimated separate intercepts for each unit. This is accomplished through the introduction of dummy variables for each unit (minus one, which avoids the dummy variable trap). This can be seen via the "α_i" in (3) above.

More intuitively, unit-specific heterogeneity refers to time-stable characteristics of the units analyzed (Cornwell and Trumbull, 1994). One county, for example, may be more conservative than the next. Alternatively, one agency may be quite different from its counterpart. As Cherry (1999) illustrated,

> "…suppose there are two criminal justice departments. Department A follows strict accounting practices that report high percentages of crimes, while department B is more lenient and reports lower percentages. Noting that certainty of sanctions is typically measured by the clearance rate, this disparity causes a problem when the reported data from the two jurisdictions are analyzed" (p. 754).

Since heterogeneity is concerned with "differences," it should also be pointed out that time periods differ from one another. The possibility exists that events occur in certain time periods that affect all of the units simultaneously. One example could be a downturn in the economy. Another example could be a terrorist incident or other highly publicized violent incident. To the extent such possibilities exist and affect the units included in the panel dataset, researchers should model them. This is accomplished by the addition of dummy variables for each time period (less one). This can be seen via the "δ_t" in (3) above.

Panel data analysts often default to unit and time dummies, and sometimes include them in their models without any attention to whether they are truly needed. How is a researcher to decide whether unit and time heterogeneity should be controlled for? A simple F-test of either the unit and time dummies (or both) suffices. A significant value suggests they should be included in the model. In most instances the F-tests *are* significant, hence the common use of unit and time dummies in panel data analysis. It is also critical to point out that there are variations on the dummy variable approach. One is the inclusion of linear trends, which will be covered shortly.

4.2 Dynamics

Panel data are rarely independent across the time dimension. Researchers expect, and routinely see, serial correlation (or temporal autocorrelation) in the data. These "dynamics" need to be modeled. Indeed, it is not uncommon for the values of a particular unit from one time period to be associated with values for the same unit from another time period (Hanushek and Jackson, 1977; Maddala, 1992). A good example concerns the public budgeting process. According to Worrall and Pratt (2003),

> "It is commonly understood that when public agencies do not spend their budgetary allotments for a particular year, their budgets can be reduced in subsequent years. Thus, the very nature of the public budgeting process ensures that an agency's budget for one year is highly associated with its budget for the previous year" (p. 89).

There are several tests available for detecting autocorrelation in panel data. Perhaps the most straightforward is the panel data analog of the standard Lagrange multiplier test. This is accomplished by estimating the OLS regression equation and then regressing the residuals on all of the independent variables and the lagged residual. If the coefficient on the lagged residual is significant, then the research can conclude autocorrelation exists. The test can be modified to handle more complex dynamic processes.

Two common methods are used to correct for autocorrelation, if it is present. One is through the introduction of one or more lags of the dependent variable. Political scientists routinely use this approach, especially in light of Beck and Katz's (1995) important study on the subject. A number of criminologists have done the same (e.g., Marvell and Moody, 1996;

Kovandzic, Sloan, and Vieraitis, 2002). The upside of this approach is twofold. One is that it expressly models autocorrelation via a coefficient that can be interpreted. Another is that it has the benefit of controlling for omitted lagged effects (Marvell and Moody, 1996).

A problem with including lagged dependent variables to control for autocorrelation is that they are probably correlated with the error term. This is particularly problematic when the time series is short (Hsiao, 1986). Also, lagging the dependent variable potentially results in lost observations (e.g., the lagged dependent variable at time 1 can no longer be treated as a dependent variable because there is no previous value of the variable to act as a (lagged) predictor). As an alternative, some researchers opt for regressions that include first-order autoregressive disturbance terms. These are arrived at by estimating ρ from β in a residual regression of $u_{it} = \beta u_{i,t-1} + \eta_{it}$ (Mundlak, 1978; Hsiao, 1986). Most statistics packages have a routine for such estimation (e.g., Stata's –xtregar– command).

It is worth digressing for a moment to further consider the issue of dynamics in the context of a short time series. When the time series is short, as we saw, controlling for autocorrelation with lagged dependent variables gobbles up degrees of freedom. Doing so also downwardly biases the coefficient on the lagged dependent variable, a problem known as Nickell bias, in reference to the individual who identified it (Nickell, 1981). Ideally, a panel data should have many time periods relative to units, but this is not always possible. Several criminologists have seen fit to ignore the issue altogether (e.g., Greenberg and West, 2001; Zhao, Scheider and Thurman, 2002).

4.3 Panel heteroskedasticity and contemporaneous correlation

Two additional problems that arise in the panel data context have been labeled panel heteroskedasticity and contemporaneous correlation (Beck and Katz, 1995). The first refers to unit-to-unit variances in the errors. To illustrate, the scale of the dependent variable, such as the crime rate, can differ markedly across units, something that could need accounting for. Franzese (2002) proposed a simple method for detecting panel heteroskedasticity. It requires regressing the absolute values of the OLS residuals on the X variable that is thought to be closely associated with the errors. In the case of panel data, the unit-specific dummy variables (minus one) are those most likely to be associated with the error term. A significant F-test for those variables is indicative of heteroskedasticity.

The second problem, contemporaneous correlation, refers to correlated errors between two or more (though not necessarily all) units at the same time. This can occur when some development in one unit is linked in some fashion to another unit. A panel data analysis of several counties from several states might face this problem if, say, an event in one state affected all or several of the counties in that state at the same time. One could simply model such effects with state (in addition to county) dummies, but the resulting models can quickly become unwieldy and result in additional losses to degrees of freedom. The test for detecting contemporaneous correlation is somewhat more complicated than the one for detecting panel heteroskedasticity, but it is described in detail in Breusch and Pagan (1980). Stata's user-written –xttest2– command is also helpful.

How is one to deal with panel heteroskedasticity and/or contemporaneous correlation? Some researchers have ignored it altogether, but one of the more significant advances from research in political science is Beck and Katz's (1995) "panel corrected standard errors" (PCSEs) approach. They proposed a relatively simple method for estimated panel data models with errors corrected for the panel data problems just highlighted. Their Monte Carlo simulations showed that PCSEs are accurate in the presence of both contemporaneous correlation and panel heteroskedasticity. Several statistics packages now contain routines for estimating models with panel corrected standard errors.

4.4 Nonstationarity

Another panel data issue concerns stationarity. Panel data must be stationary. In formal terms, this means their means, variances, and autocovariances (at various lags) remain constant across all time points. An augmented Dickey-Fuller test can be performed to detect stationarity/nonstationarity. This is accomplished by regressing the first-differenced dependent variable on the one-period lag of the dependent variable and lags of the first-differenced dependent variables (Enders, 1995).

4.5 Trends

As pointed out already, the typical fixed effect regression adds unit and year dummies (this is known as a "full fixed effects" or "two-way fixed effects" model). An extension of this approach used by some researchers is to include separate deterministic trends for each unit of analysis. These trends (coded from one to T for each unit) control for trends in each unit that depart from the annual shocks captured by the year dummies (see, e.g., Black and Nagin, 1998; Marvell and Moody, 1996; Worrall, 2005). The trends amount to proxies for factors that make values of the dependent variable in one unit grow more or less than the other units as a whole. That is, they model departures from the norm. The problem, though, is that they can be highly collinear with other variables that also trend upward.

As a twist on the trends-for-each-unit approach, some researchers have used a single trend variable (not tied to any specific unit) to model the incapacitative effect of crime control legislation. A common problem for policy researchers has been to separate deterrent effects (the individual chooses not to commit crime, for example, because of fear of punishment) from incapacitative effects (the individual is rendered—usually physically—incapable of committing crime, e.g., by imprisonment) of such policies on crime. An early approach, taken by Kessler and Levitt (1999), involved examining the crime rate immediately following the passage of a new policy. Since incapacitation is unlikely to influence crime in the short run (since actual capture and imprisonment would presumably not be immediate, but would take some time to implement) any resulting decline in crime would have to result from deterrence (*fear* of capture and imprisonment as a result of the new policy).

A more recent approach, taken by Marvell and Moody (2001), was to include a linear trend starting at the time a new policy is passed (in addition to other variables). In their view, an incapacitation trend variable "…assumes that, in the absence of the law, very few defendants would have escaped prison sentences, so that the incapacitation effect grows over time" (p. 103). Researchers can get even more sophisticated with the use of trend variables. Various combinations of interactions and even nonlinear trends can be modeled in an effort to capture interesting and complicated relationships.

5 A practical example: welfare spending and crime

To illustrate several of the issues raised thus far, let us revisit the author's recent study on the relationship between welfare spending and crime (Worrall, 2005). The data for that study, which are also used here, consisted of yearly observations from 1990 to 1998 for all 58 counties in California. We will more or less replicate the results from that study here, but with an eye toward understanding, first, what happens when data are simply pooled ignoring the distinction between the time series and cross-sectional components. Then we will progressively account for the realities of panel data by first introducing fixed effects, then adding trends, and finally adding panel corrected standard errors. Dynamics will be dealt with via a lagged dependent variable, and it should also be pointed out that the data are stationary according to the augmented Dickey-Fuller test. The analyses also will be based on logged variables to minimize the effects of outliers (e.g., Marvell and Moody, 1996).

The analyses reported below are more than illustrative. They offer a critical take on a wealth of previous research aimed at detecting the relationship between welfare spending and crime (see Worrall, 2005, for a review). Most such studies have employed cross-sectional designs and, in doing so, have found an inverse relationship between welfare spending and crime. Panel data, however, permit researchers to control for various unobserved time-stable (and time-period-specific) effects. But pooling alone does not ensure replication of results from OLS performed on cross-sectional data. As will become clear, when the data for the present analysis are pooled, a previously inverse relationship became positive. That is, there was more crime in areas characterized by higher levels of welfare spending. Why? Panel heteroskedasticity is a likely candidate. Omitted variable bias is another (for additional explanations see Kennedy, 2002). Panel data models help get around these very problems. As will become clear, fixed effects modeling causes the inverse relationship to reappear.

5.1 Variables

The dependent variable in the analyses reported below consists of the rates per 100,000 people of homicide, robbery, assault, burglary, and larceny. There were no homicides in 80 county/years, so a value of .5 was added to the homicide variable. The measure of welfare spending is the cost-of-living adjusted Aid to Families with Dependent Children (AFDC) annual payment per recipient. A similar measure has been used in a number of similar studies on this subject (e.g., DeFronzo, 1983; DeFronzo and Hannon, 1998). The independent variables are population mobility (Mobility), the poverty level (Poverty), the percentage of single mother households (Mothers), population density (Density), percent Black (Black), percentage young males between the ages of 13 and 17 (Male 1), and the percentage of young males between the ages of 18 and 25 (Male 2) (see Worrall, 2005, for additional details as well as summary statistics).

5.2 Uncorrected OLS model

The results reported in Table 15.1 are from a pooled OLS regression model. The results are equivalent to those of a basic OLS model with the only difference being repeated observations on each of the units. The first coefficients reported in Table 15.1 are the lagged dependent variables (to control for autocorrelation). All are significant and positive, as expected. Welfare spending, however, is not inversely associated with crime in the pooled model. Additionally, the relationship is significant for homicide, assault, burglary, and larceny. This stands in contrast to the bulk of previous research on the subject (see, e.g., Worrall, 2005). Most past studies show less crime in areas characterized by high welfare spending, but pooling the data as we have here, without regard to the fact that there are repeated observations on each unit, is the likely explanation for this finding. In short, we cannot make much of the results reported in Table 15.1 because the model is incorrectly specified, but it does at least provide a base for comparison. The next step is to add fixed effects.

5.3 Two-way fixed effects model

Table 15.2 reports the results of models with county and year fixed effects (i.e., dummy variables) added to the models. Two observations are apparent. First, all but one of the AFDC coefficients shifted their signs, consistent with most previous studies. What's more, all but one lost

Table 15.1 Uncorrected OLS model

	Homicide	*Robbery*	*Assault*	*Burglary*	*Larceny*
Lagged dep. var.	0.467 (12.02)**	0.591 (16.64)**	0.754 (25.83)**	0.812 (28.79)**	0.943 (57.40)**
AFDC	0.671 (2.96)**	0.281 (0.97)	0.208 (2.12)*	0.246 (4.10)**	0.224 (5.37)**
Mobility	−0.309 (1.13)	−1.062 (2.85)**	−0.111 (0.91)	−0.123 (1.55)	−0.009 (0.17)
Poverty	0.424 (3.34)**	−0.089 (0.55)	0.085 (1.57)	0.052 (1.48)	0.001 (0.04)
Mother	−0.199 (0.76)	1.752 (4.65)**	0.256 (2.18)*	0.134 (1.91)	0.045 (0.86)
Density	0.007 (0.25)	0.166 (4.40)**	−0.010 (0.84)	−0.007 (0.92)	−0.002 (0.45)
Black	0.041 (1.09)	0.030 (0.61)	−0.028 (1.72)	−0.014 (1.36)	−0.008 (1.12)
Male 1	0.541 (1.72)	−0.302 (0.73)	−0.192 (1.39)	0.032 (0.37)	−0.001 (0.02)
Male 2	−0.139 (1.00)	−0.466 (2.50)*	−0.010 (0.16)	−0.063 (1.62)	0.009 (0.33)
Constant	−9.128 (3.60)	0.124 (0.04)	−2.871 (2.73)	−2.213 (3.39)	−1.825 (4.01)
Observations	513	513	513	513	513
R-squared	0.39	0.81	0.67	0.78	0.91

Notes: Absolute value of t statistics in parentheses.
* significant at 5%;
** significant at 1%. All F-statistics were significant at the $p = .001$ level. All variables expressed as per capita logarithms. County and year dummy output suppressed.

their significance (compared to Table 15.1). It appears there may be a significant inverse relationship between robbery and welfare spending. Importantly, though, the models reported in Table 15.2 make no other corrections for the panel data structure besides (1) adding fixed effects and (2) controlling for autocorrelation. Other problems, such as panel heteroskedasticity, contemporaneous correlation, and trends, are ignored. The following sections consider the influence of such problems.

5.4 PCSEs

Table 15.3 extends the models reported in the previous section but adds panel corrected standard errors, to correct for panel heteroskedasticity and contemporaneous correlation (which, incidentally, were present in the data, based on the tests reported earlier). Even after PCSEs were added to the models (via Stata's –xtpcse– command), however the AFDC coefficient for robbery remained negative and significant. It would appear, then, that welfare spending may

Table 15.2 Fixed effects added

	Homicide	*Robbery*	*Assault*	*Burglary*	*Larceny*
Lagged dep. var.	−0.090	−0.129	0.416	0.363	0.462
	(1.93)	(2.74)**	(10.00)**	(8.26)**	(10.74)**
AFDC	−0.688	−3.588	−0.141	0.234	−0.193
	(0.83)	(3.43)**	(0.35)	(1.02)	(1.17)
Mobility	−0.099	0.259	0.784	0.000	0.521
	(0.10)	(0.22)	(1.68)	(0.00)	(2.74)**
Poverty	−2.291	4.532	−0.010	0.192	0.225
	(3.04)**	(4.76)**	(0.03)	(0.92)	(1.49)
Mother	0.476	−1.771	0.138	0.457	−0.111
	(0.71)	(2.08)*	(0.42)	(2.44)*	(0.83)
Density	−0.672	−1.855	−1.180	−0.291	0.105
	(0.64)	(1.40)	(2.30)*	(0.99)	(0.51)
Black	−0.325	−0.292	−0.302	−0.659	−0.117
	(0.88)	(0.63)	(1.69)	(6.41)**	(1.58)
Male 1	0.129	0.346	−0.925	−0.329	0.119
	(0.26)	(0.55)	(3.67)**	(2.33)*	(1.20)
Male 2	−0.087	0.657	0.189	0.324	0.029
	(0.37)	(2.24)*	(1.68)	(4.91)**	(0.62)
Constant	5.184	20.817	0.619	−4.853	−4.226
	(0.42)	(1.35)	(0.10)	(1.43)	(1.79)
Observations	513	513	513	513	513
R-squared	0.64	0.90	0.77	0.86	0.94

Notes: Absolute value of t statistics in parentheses.
* significant at 5%;
** significant at 1%. All F-statistics were significant at the $p = .001$ level. All variables expressed as per capita logarithms. County and year dummy output suppressed.

be associated with reductions in serious crime. This finding stands in contrast to the author's 2005 study (Worrall, 2005), for two reasons. First, that study controlled for autocorrelation by estimating AR(1) disturbance terms in lieu of lagging the dependent variable. Second, the 2005 study included county-specific trends, the logic of which was discussed earlier. If we add such trends, as the models reported in the following section do, the effect of welfare spending on crime disappears altogether.

5.5 Thinking about trends

Table 15.4 extends the models from the previous section but with the addition of separate county-specific trends (to model departures from statewide trends). As can be seen in Table 15.4, there appears to be no effect of welfare spending on serious crime. This is what the author reported in Worrall (2005). AFDC, of course, is not the only measure of welfare spending. When additional measures are included, as the author did in the 2005

Table 15.3 PCSEs added

	Homicide	*Robbery*	*Assault*	*Burglary*	*Larceny*
Lagged dep. var.	−0.090 (0.75)	−0.129 (0.71)	0.416 (4.01)**	0.363 (3.11)**	0.462 (3.71)**
AFDC	−0.688 (0.75)	−3.588 (3.37)**	−0.141 (0.43)	0.234 (1.10)	−0.193 (1.23)
Mobility	−0.099 (0.12)	0.259 (0.38)	0.784 (2.18)*	0.000 (0.00)	0.521 (3.51)**
Poverty	−2.291 (3.58)**	4.532 (3.13)**	−0.010 (0.03)	0.192 (0.93)	0.225 (1.98)*
Mother	0.476 (0.92)	−1.771 (3.01)**	0.138 (0.31)	0.457 (1.77)	−0.111 (0.90)
Density	−0.672 (0.68)	−1.855 (2.00)*	−1.180 (1.97)*	−0.291 (0.97)	0.105 (0.44)
Black	−0.325 (1.11)	−0.292 (0.81)	−0.302 (1.19)	−0.659 (3.39)**	−0.117 (0.90)
Male 1	0.129 (0.33)	0.346 (0.49)	−0.925 (2.18)*	−0.329 (1.27)	0.119 (0.58)
Male 2	−0.087 (0.39)	0.657 (1.75)	0.189 (1.46)	0.324 (5.51)**	0.029 (0.47)
Constant	5.184 (0.52)	20.817 (1.57)	0.619 (0.09)	−4.853 (1.52)	−4.226 (1.57)
Observations	513	513	513	513	513

Notes: Absolute value of t statistics in parentheses.
* significant at 5%;
** significant at 1%. All F-statistics were significant at the $p = .001$ level. All variables expressed as per capita logarithms. County and year dummy output suppressed.

study, even more evidence suggests that welfare spending has little to no effect on serious crime. The author also checked the robustness of the findings through other means (such as by excluding various combinations of variables) and even by using a linear specification, each of which pointed to the independence of welfare spending and crime. This finding is important, once again, because it flies in the face of scores of previous studies. Panel data provide an interesting opportunity to model heterogeneity, and it appears that once such characteristics are modeled, and once an appropriate specification is used, a more complete story can be told. As the author concluded in the 2005 study,

> "The finding suggests that the relationship between welfare spending and macro-level crime rates is practically nonexistent. This is likely due to statistical controls for unobserved heterogeneity and the introduction of dummy variables for each year (minus one)" (Worrall, 2005, p. 365).

Researchers often specify their panel data models with little more than lip service to the issues presented when time series and cross-section data are pooled. The steps taken

Table 15.4 Trends added

	Homicide	*Robbery*	*Assault*	*Burglary*	*Larceny*
Lagged dep. var.	−0.226 (1.84)	−0.319 (1.93)	0.119 (0.97)	0.014 (0.12)	0.110 (0.79)
AFDC	−0.859 (0.88)	−1.428 (1.36)	0.051 (0.09)	0.338 (1.22)	−0.205 (0.93)
Mobility	3.995 (0.39)	−5.967 (0.41)	10.060 (1.19)	6.397 (1.83)	−1.903 (0.51)
Poverty	10.282 (0.60)	−34.367 (1.00)	−4.532 (0.66)	−1.060 (0.20)	2.606 (1.29)
Mother	8.263 (1.75)	0.490 (0.13)	−6.250 (1.81)	−3.581 (1.36)	−1.440 (0.95)
Density	−2.594 (1.88)	0.310 (0.14)	−1.332 (1.66)	−0.130 (0.31)	−0.465 (1.36)
Black	0.123 (0.22)	0.292 (0.83)	0.235 (0.50)	−0.355 (1.12)	−0.395 (1.36)
Male 1	0.815 (0.89)	1.085 (0.90)	−0.235 (0.43)	−0.118 (0.34)	0.140 (0.60)
Male 2	−2.230 (1.66)	−3.346 (1.82)	−0.039 (0.06)	0.228 (0.69)	0.677 (2.44)*
Constant	−4.859 (0.10)	75.260 (0.86)	−49.732 (1.30)	−35.539 (1.65)	2.991 (0.19)
Observations	513	513	513	513	513

Notes: Absolute value of t statistics in parentheses.
* significant at 5%;
** significant at 1%. All F-statistics were significant at the $p = .001$ level. All variables expressed as per capita logarithms. County and year dummy output suppressed.

here were partly for illustration, but they were also arguably necessary. First, the basic OLS model returned some strange results, several positive and significant relationships between welfare spending and crime. Fixed effects cleared that up. But the data also displayed panel heteroskedasticity and contemporaneous correlation, hence the addition of panel corrected standard errors. Finally, models were estimated with county-specific trends, which eliminated the significant relationship between welfare spending and robbery. If they were *not* necessary, then it still appears that welfare spending has almost no effect on serious crime (save for robbery). If people commit crime for sustenance and welfare payments dissuade such behavior (DeFronzo, 1983), then one would expect reductions in burglary and larceny as well, but the coefficients in those models were never significant, regardless of the specification.

6 Simple extensions

The previous section provided a simple illustration of pooling time series and cross-section data and how coefficients can change depending on the extent to which specification issues are addressed. Controlling for heterogeneity and adding unit-specific-time trends appeared to have the most pronounced effect. But there are still other steps researchers can take to extend panel data models such that they tell a more complete story. This section briefly introduces such extensions, but not by way of example. Instead, what is presented here is a simple conceptual overview (with some how-to) of three extensions: (1) dealing with simultaneity; (2) estimating the effects of events; and (3) exploring unit-specific effects. Each step has been taken in a number of criminological studies. The intent here is to introduce them and describe how and why one might want to pursue them.

6.1 Dealing with simultaneity

Returning briefly to the welfare spending–crime discussion in the previous section, Piven and Cloward (1971) have argued that welfare programs have at their core a sinister motive to control the poor. If crime is committed disproportionately by poor persons (as it appears to be), then logic suggests governments may alter welfare benefits in an effort to keep crime, and therefore the poor, in check. Put another way, governments may use crime as a consideration in setting welfare benefits. This is but one variation of the so-called simultaneity problem, a two-way relationship between the independent and dependent variables.

Welfare spending may reduce crime, but crime may affect welfare spending. If one is not convinced of the connection here, then consider police spending and crime. It is much more logical to conclude that police levels, either officers per citizen or police spending per citizen, may reduce crime. It is also quite logical to conclude that crime may cause units of government to alter police levels. Getting to the bottom of such two-way relationships is of critical importance.

OLS regression provides researchers with only one basic method of addressing simultaneity: instrumental variable regression as implemented in two-stage least squares. It is just as viable in the panel data context and has been used with great frequency, even in the police levels crime literature (for recent examples, see General Accounting Office, 2005; Evans and Owens, 2005). Another approach, only available in the panel data or time series context, is the Granger causality test (Granger, 1969; Pindyck and Rubinfeld, 1991, pp. 216–219). The Granger causality test has been used by a number of criminologists, including Marvell and Moody (1996) and Kovandzic et al. (2002). The test involves two separate autoregressive analyses. Let us consider the example of police levels and crime.

The first step is to regress crime on lags of itself and lags of police levels (along with pertinent controls). The lagged variables are dropped if they are not significant. If lags of police levels are significant (as determined by an F-test), police levels "Granger cause" crime. The second test is the opposite, with police levels serving as the dependent variable. The model includes lags of police levels and crime. If lagged crime is significant in such a model, crime "Granger causes" police. There is one key limitation of the Granger causality test, notably its inability to detect instantaneous effects. That is, because the test relies on lags, in our example it would be impossible to determine whether police levels affect crime (or vice-versa) in the same time period. However, the lack of a lagged impact implies the lack of a current impact because lagged crime is likely to be correlated with current crime through serial correlation.

6.2 Estimating the effects of events

Panel data models have long been considered some of the best designs for the study of causation next to a purely random experiment. For example, Campbell and Stanley (1967, pp. 55–57) refer to panel models as "excellent quasi-experimental design[s], perhaps the best of the more feasible designs." Lempert (1966, pp. 130–131) stated that panel designs are research designs "*par excellence*." Still other researchers have argued that panel techniques are well-suited to causal analysis (e.g., Stimson, 1985; Hsiao, 1986). One of the reasons panel data designs are regarded in this manner is because they can be used to estimate the effects of events, not unlike a classic experiment. Treatment (units which experience the event) and control (units which do *not* experience the event) can be included in the same analysis.

Estimating the effects of events with panel data is quite straightforward. The researcher has at least three options at his/her disposal. The first is to measure the presence or absence of an event with a 1/0 dummy variable. The event could be something that unfolds during the time span covered by the data, or it could be the presence or absence of some condition (such as a law). The only limitation to this approach is that it cannot be employed when (1) other time invariant variables are included in the model and (2) every unit experiences the event at the same time (see Allison, 1994). The second approach is to measure the event in levels. For example, Zhao et al. (2002) used panel data to assess the effects of COPS spending on crime. Rather than a 1/0 dichotomy for the presence or absence of a COPS grant, their main variable was an estimate of the amount of grant funds. The third approach is to adopt more sophisticated coding schemes, particularly in cases where researchers are interested in more than the presence or absence of some event or condition. Various linear and nonlinear coding schemes are possible (see Allison, 1994, pp. 185–187).

6.3 Exploring unit-specific effects

Panel data analysts are sometimes guilty of something known as the "geographic aggregation assumption" (Black and Nagin, 1998, p. 213). This amounts to assuming that the parameters are constant across units. For example, a researcher may estimate a panel data model to measure the effect of an event, such as a new law, on crime. The model will return one coefficient for all the units/time periods. The assumption that an intervention has the same effect on crime across the board (as indicated by a single coefficient) can be quite perilous. Pesaran and Smith (1995) found, for example, that pooling panels (e.g., estimating a single coefficient that applies across all units and time periods) can result in significant bias: "[g]iven the prevalence of aggregation and pooling in applied work, these results are of some importance, and indicate that the common assumption of homogeneity in dynamic models is far from innocuous" (p. 102). Black and Nagin (1998) used this notion to attack Lott and Mustard's (1997) "more guns, less crime" thesis.

How do researchers deal with this problem, if it exists? By interacting unit dummies with the predictor variable of interest. This yields separate estimates for each unit.

Black and Nagin interacted a Lott and Mustard's "right-to-carry" variable (1,0 variable denoting the presence or absence of a right-to-carry law) with state dummy variables to come up with different estimates for each state. They also experimented with a variation on this approach to test whether the laws had the same effect on crime across different time periods. Lott and Mustard (1997) found a significant inverse relationship between right-to-carry laws and several indicators of serious crime. When state-specific effects were explored, however, Black and Nagin (1998, pp. 213–14) concluded: "…we strongly reject the Lott and Mustard model's assumption of a uniform impact across states…The estimates are disparate. Murders decline in Florida but increase in West Virginia. Assaults fall in Maine but increase in Pennsylvania…"

7 Summary and additional readings

The purpose of this chapter has been to offer a simple introduction to the pooling of time series and cross-section data (both of which yield a "panel" dataset). It began by highlighting the various issues in panel data and then covered means of testing and controlling for them. This was followed by a practical example of the links between welfare spending and serious crime. The last few sections explored extensions, such as dealing with simultaneity, modeling events, and moving beyond the homogeneous parameters assumption. Sources on the analysis of panel data accessible to individuals with minimal statistical background include Allison (1994), Cherry (1999), Lott and Mustard (1997), Stimson (1985), and Worrall and Pratt (2004). More detailed and technical treatments include Wooldridge (2001), Hsiao (1986; 2003), Arellano (2003), and Frees (2004).

References

Allison, P.D. (1994). Using panel data to estimate the effects of events. *Sociological Methods and Research*, 23: 174–199.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford: Oxford University Press.

Beck, N. and Katz, J. N. (1995a). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, 89: 634–74.

Beck, N. and Katz, J. N. (1995b). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, 89: 4–647.

Beck, N. and Katz, J. N. (2004). Time-series—cross-section issues: Dynamics. Unpublished manuscript.

Berk, R. A. (2004). *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Black, D. A. and Nagin, D. S. (1998). Do right-to-carry laws deter violent crime? *Journal of Legal Studies*, 27: 209–219.

Breusch, T. and Pagan, A. (1980). The LM test and its applications to model specification in econometrics. *Review of Economic Studies*, 47: 239–254.

Campbell, D. T. and Stanley, J. C. (1967). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cherry, T. L. (1999). Unobserved heterogeneity bias when estimating the economic model of crime. *Applied Economic Letters*, 6: 753–757.

Cornwell, C. and Trumbull, W. N. (1994). Estimating the economic model of crime with panel data. *Review of Economics and Statistics*, 76: 360–366.

DeFronzo, J. (1983). Economic assistance to impoverished Americans: Relationship to the incidence of crime. *Criminology*, 21: 119–136.

DeFronzo, J. and Hannon, L. (1998). Welfare assistance levels and homicide. *Homicide Studies*, 2: 31–45.

Enders, W. (1995). *Applied Econometric Time Series*. New York: Wiley.

Evans, W. N. and Owens, E. (2005). Flypaper COPS. College Park, MD: University Maryland, Department of Economics.

Franzese, R. J. (2002). *Macroeconomic Policies of Developed Democracies*. Cambridge, UK: Cambridge University Press.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge, UK: Cambridge University Press.

General Accounting Office (2005). *Community Policing Grants: COPS Grants were a Modest Contributor to Declines in Crime in the 1990s*. Washington, DC: General Accounting Office.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37: 424–438.

Greenberg, F. D. and West, V. (2001). State prison populations and their growth, 1971–1991. *Criminology*, 3: 615–654.

Hanushek, E. A. and Jackson, J. E. (1977). *Statistical Modeling for Social Scientists*. San Diego, CA: Academic Press.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46: 1251–1271.

Hsiao, C. (1986). *Analysis of Panel Data*. New York: Cambridge University Press.

Hsiao, C. (2003). *Analysis of Panel Data*, 2nd edn. New York: Cambridge University Press.

Kennedy, P. (2002). Oh no! I got the wrong sign! What should I do? Unpublished manuscript. Simon Fraser University, British Columbia.

Kessler, D. and Levitt, S. D. (1999). Using sentence enhancements to distinguish between deterrence and incapacitation. *Journal of Law and Economics*, 42: 343–363.

Kovandzic, T., Sloan, J. J. III and Vieraitis, L. M. (2002). Unintended consequences of politically popular sentencing policy: The homicide promoting effects of "three strikes" in US cities

(1980–1999). *Criminology and Public Policy*, 1: 399–424.

Kristensen, I. P. and Wawro, G. (2003). Lagging the dog? The robustness of panel corrected standard errors in the presence of serial correlation and observation specific effects. Presented at the Annual Meeting of the Society for Political Methodology, University of Minnesota.

Lempert, R. (1966). Strategies of research design in the legal impact study: The control of plausible rival hypotheses. *Law and Society Review*, 1: 111–132.

Lott, J. R. and Mustard, D. B. (1997). Crime, deterrence, and right-to-carry concealed handguns. *Journal of Legal Studies*, 26: 1–68.

Maddala, G. S. (1992). *Introduction to Econometrics*. New York: Macmillan.

Marvell, T. B. and Moody, C. E. (1996). Specification problems, police levels, and crime. *Criminology*, 34: 609–643.

Marvell, T. B. and Moody, C. E. (2001). The lethal effects of three-strikes laws. *Journal of Legal Studies*, 30: 89–106.

Mundlak, Y. (1978). On the pooling of time-series and cross-section data. *Econometrica*, 4: 69–85.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49: 1417–1426.

Pesaran, M. H. and Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68: 79–113.

Pindyck, R. S. and Rubinfeld, D. L. (1991). *Econometric Models and Economic Forecasts*. New York: McGraw-Hill.

Piven, F. F. and Cloward, R. A. (1971). *Regulating the Poor: The Functions of Public Welfare*. New York: Vintage.

Stimson, J. (1985). Regression in space and time: A statistical essay. *American Journal of Political Science*, 29: 914–947.

Wooldridge, M. M. (2001). *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT Press.

Worrall, J. L. (2005). Reconsidering the relationship between welfare spending and serious crime: A panel data analysis with implications for social support theory. *Justice Quarterly*, 22: 364–391.

Worrall, J. L. and Pratt, T. C. (2004). On the consequences of ignoring unobserved heterogeneity when estimating macro-level models of crime. *Social Science Research*, 33: 79–105.

Zhao, J., Scheider, M. C. and Thurman, Q. (2002). Funding community policing to reduce crime: Have COPS grants made a difference? *Criminology and Public Policy*, 2: 7–32.

Chapter 16

Dynamic models and cross-sectional data: the consequences of dynamic misspecification

Ronald Schoenberg

1 Introduction

Most social variables are inherently imbedded in time and, therefore, are generated by a dynamic process of some kind. This paper explores the consequences of applying a static model to such variables. We find that unbiased estimates of the underlying dynamic parameters through the application of a static model to a cross-section of data are possible only if the underlying dynamic process is "nonergodic," i.e., that the process is a function of the initial conditions. A wide-sense stationary dynamic model that is not a function of the initial state of the system is proposed here, and we find that the application of static models to such processes is not successful, in particular we find that the static estimates are attenuated or inflated versions of their dynamic counterparts.

In the analysis of cross-sectional data, which constitutes most of the data available for analysis in sociology, some form of a linear structural model is very often proposed and appropriately so. Theoretical specification is often unable to rise above the simple assertion that, when X increases Y increases or decreases in proportion (actually such a claim may not be so simple when imbedded within an array of such claims with regard to a large set of variables). When this is so, the representation of such assertions in linear structural equation models is correct, provided as well, of course, that assumptions about the behavior of excluded variables and in some cases about the distribution of the variables themselves in the population are plausible. And the estimated parameters of these models have immediate substantive interpretation relative to the hypothesized structure of relations of the variables in the model. That is, given the adequacy of the model, we may expect that a given increase in one variable of the model will have the expected consequences for the other variables in the model which that variable is hypothesized to affect. A clearly defined relationship exists, then, between the model and the "reality" out there.

Many other models proposed for cross-sectional data, however, do not represent directly the claims of the theory but, rather, are surrogate models taking the place of dynamic models. The source of the claims regarding the

Reprinted from *Social Science Research* 2:133–144, 1977, with permission from Elsevier

underlying process that has generated the data may specify that the variables are dynamically related, but the form of the collected data—the product of, say, a one-shot cross-sectional study—may preclude the specification of a dynamic model. In these cases, investigators commonly propose a linear structural equation model of some kind and proceed to estimate and interpret parameters while neglecting the dynamic nature of the underlying process.

Such misspecification may be more common than supposed, since nonexperimental sociological variables are inherently time-dependent. Time dependency doesn't necessarily imply that a dynamic model must be specified, but it does mean that the behavior of the variables across time must be scrutinized and that the investigator must determine whether or not this behavior is such that it precludes unbiased efficient estimates of the structural parameters. In the following sections, we shall explore various time-dependent configurations and we shall determine precisely the consequences of each on the application of static models to underlying dynamic processes.

Similar investigations have been conducted by other authors. The results here are generalizations of some discussion of this issue by Kuh (1959). Furthermore, some of the hazards of specifying a static model have been pointed out by Heise (1975) using a simplified systems analysis formulation which did not include residual error terms. Related issues surrounding the estimation and specification of models that incorporate both cross-sectional and time-series features appear in Balestra and Nerlove (1966), Young (1972), and Theil and Goldberger (1961).

2 General dynamic linear structural equation model

Let Z_t be a vector of p variables at time t, A a $p \times p$ matrix of coefficients, and Υ_t be a vector of residual variables. Let

$$Z_t = AZ_{t-1} + \Upsilon_t. \tag{1}$$

Furthermore, let $E(\Upsilon_t) = 0$ and $E(\Upsilon_t \Upsilon_s') = \delta_{ts}\Xi$, where $\delta_{ts} = \begin{cases} 1 & t = s \\ 0 & t \neq s \end{cases}$ i.e., the elements of Υ_t are not autocorrelated and have the variance–covariance matrix Ξ. Equation (1) is a wide-sense stationary linear first-order difference equation, provided that the coefficient matrix A is non-singular and that its characteristic roots are less than one in modulus. Then,

$$Z_t = \sum_{r=0}^{\infty} A^r \Upsilon_{t-r} \tag{2}$$

is a solution of equation (1) which is independent of the initial state of the system and is stable (Miller, 1968, Chapter 4). If the characteristic roots are greater than 1 in modulus, the system will be unstable and divergent, which is clearly a property of only shortlived processes and, for which, specific models must be proposed rather than the general type discussed here (certainly studying such a process cross-sectionally would be hopeless). If the roots are zero or equal to one, then the process is a function of the initial conditions. The analysis of such models cross-sectionally is possible; however, a substantive argument that would support a dynamic model of a large-scale social process that is strongly dependent on the initial state of the system must be carefully made.

For our purposes here, we wish to derive an expression for the variance–covariance matrix of Z at time t in terms of the coefficient matrix A and the variance–covariance matrix Ξ of the residual variables. From equation (2),

$$E(Z_t Z_{t+k}') = \sum_{r=0}^{\infty} A^r \Xi A'^{(r+k)} \tag{3}$$

$$E(Z_{t+k} Z_t') = \sum_{r=0}^{\infty} A^{(r+k)} \Xi A' \tag{4}$$

From equation (1), $\Xi = (Z_t - AZ_{t-1})\,(Z_t - AZ_{t-1})'$ $= Z_t Z_t' \ - AZ_{t-1}Z_t' - Z_t Z_{t-1}' A' + AZ_{t-1}Z_{t-1}'A'$; but then, from equations (3) and (4), we get

$E(Z_tZ'_{t-1}) = AE(Z_tZ'_t) = A\Sigma$ and $E(Z_{t-1}Z'_t) = E(Z_tZ'_t)A' = \Sigma A'$ and, therefore,

$$E(\Xi) = \Sigma - A\Sigma A' - A\Sigma A' + A\Sigma A' = \Sigma - A\Sigma A' \quad (5)$$

3 Quasi-dynamic model

A very commonly specified linear static or cross-sectional model is the multiple regression model: $y = \Gamma X + z$, where y is a dependent variable, X is a vector of independent variables, Γ is a vector of regression coefficients, and z is a residual variable.[1] The well-known estimate of Γ is $\hat{\Gamma} = yX'(XX')^{-1} = \Sigma_{yx}\Sigma_{xx}^{-1}$. Suppose that the variances and covariances in Σ were generated by an analogous dynamic structural model,

$$y_t = BX_{t-1} + u_t \quad (6)$$

$$X_t = X_{t-1} + V_t \quad (7)$$

or in the form of equation (1),

$$\begin{bmatrix} y_t \\ X_t \end{bmatrix} = \begin{bmatrix} 0 & \mathrm{B} \\ 0 & \mathrm{I} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ V_t \end{bmatrix}$$

Equation (7) implies that X is fixed in an expected sense, i.e., $E(X_t)$ and $E(X_tX'_t)$ are the same for all t. We can see immediately that the coefficient matrix of this model is singular. The variance–covariance matrix in this case will be a function of the initial state of the system. We will, despite this obstacle, be able to calculate the estimate of Γ in terms of the parameters of the dynamic model if we assume that the initial values are random and uncorrelated with residual error terms. The solution of equation (7) is

$$X_t = X_0 + \sum_{r=0}^{t} V_r$$

and substituting into equation (6), we get as its solution,

$$y_t = BX_0 + B\sum_{r=0}^{t-1} V_r + u_t$$

The population covariance matrix of y_t and x_t is $\Sigma_{yx} = E(y_tX'_t) = BX_0X'_0$, since $E(V_t) = E(u_t) = E(X_0V'_t) = E(X_0u'_t) = E(u_tV'_t) = 0$. Furthermore, $\Sigma_{xx} = E(X_tX'_t) = X_0X'_0$ and, therefore, $\hat{\Gamma} = BX_0X'_0(X_0X'_0)^{-1} = B$.

Thus, we see that the parameters of the multiple regression model are unbiased estimates of the parameters of the analogous dynamic model.

4 Autocorrelated model

In the previous section, we have seen that the static multiple regression model has no strictly analogous dynamic counterpart, at least in the sense of dynamic as developed here, a stable ergodic stationary process that is not a function of the initial conditions. However, by making some assumptions about the initial conditions, i.e., that they are uncorrelated with future disturbances, we found that unbiased estimates of the underlying nonergodic dynamic parameters resulted from the application of the static model to the process. In this section, we shall see that

[1] To simplify the present discussion, issues with regard to means, intercepts, and sampling variability will be ignored. The dynamic models discussed here, the linear difference equations, are stationary and, therefore, no loss of generality is created by assuming that that stationary point is zero. That the measurement of variables in nondynamic structural models from their respective means does not endanger generality is well known. Issues of sampling variability are not relevant to the discussion in this paper and are ignored by the presumption that we are studying a population. Those interested may refer to Goldberger (1964, p. 142) who thoroughly covers sampling problems in univariate and bivariate dynamic models. Furthermore, all variable matrices in this paper have observation indices suppressed. Thus, all variable matrices (as opposed to coefficient matrices and variance–covariance matrices) which are here considered to be vectors, are implicitly rectangular matrices of N observations from some population on the appropriate number of variables, usually indefinite. For example, Z_t in equation (1) is implicitly a $p \times N$ matrix of observations.

a static model with autocorrelated endogenous residuals also does not have an ergodic dynamic counterpart, i.e., the system is a function of the initial conditions and that, given some assumptions about the initial conditions, the properly specified static model yields unbiased, though inefficient, estimates.

First, let

$$y_t = BX_{t-1} + u_t \tag{8}$$

$$X_t = X_{t-1} + V_t \tag{9}$$

where

$$u_t = \phi u_{t-1} + u_t^* \tag{10}$$

where u_t^* and V_t have the desirable properties of being serially and mutually uncorrelated. Substituting equation (10) into equation (8), we get

$$y_t = BX_{t-1} + \phi y_{t-1} - \phi BX_{t-1} + u_t^* \tag{11}$$

Lagging equation (8) one time period and multiplying by ϕ, $\phi y_{t-1} = \phi BX_{t-2} + \phi u_{t-1}$ and substituting into equation (11):

$$y_t = \phi y_{t-1} + BX_{t-1} - \phi BX_{t-2} + u_t^* \tag{12}$$

Since the exogenous residuals are not autocorrelated,

$$X_t = X_0 + \sum_{r=0}^{t} V_r \tag{13}$$

Substituting equation (13) into equation (12),

$$y_t = (BX_0 + \phi y_{t-1}) - \left(\phi BX_0 + 2\sum_{r=0}^{t-1} V_r - V_t + u_t^*\right)$$

The cross-sectional regression estimates are $\hat{\Gamma} = \Sigma_{yx}\Sigma_{xx}^{-1} = y_t X_t'(X_t X_t')^{-1}$. But

$$E(y_t X_t') = \left(BX_0 + \phi y_{t-1} - \phi BX_0 + 2\sum_{r=0}^{t-1} V_r - V_t + u_t^*\right)\left(X_0' + \sum_{r=0}^{t} V_r'\right) = BX_0X_0' + \phi y_{t-1}X_0' - \phi BX_0X_0'$$

assuming that $E(V_s V_t', t \neq s) = 0$, and $E(u_t^* V_t') = E(u_t^* X_0') = E(y_t V_t') = 0$

Since

$$E(X_t X_t') = \left(X_0 + \sum_{r=0}^{t} V_r\right)\left(X_0' + \sum_{r=0}^{t} V_r'\right) = X_0 X_0'$$

then

$$\begin{aligned}\hat{\Gamma} &= (BX_0X_0' + \phi y_{t-1}X_0' - \phi BX_0X_0')(X_0X_0')^{-1}\\ &= B + \phi y_{t-1}X_0'(X_0X_0')^{-1} - \phi B\\ &= B + \phi\hat{\Gamma} - \phi B\end{aligned}$$

since $E(y_{t-1}X_0') = E(y_t X_0')$. Solving for $\hat{\Gamma}$, $\hat{\Gamma} = B$. Thus, we see that the cross-sectional estimates of the regression coefficients are unbiased estimates of the analogous dynamic coefficients for the case in which the endogenous residuals only are autocorrelated. Other authors that have made this point (Goldberger, 1964; Hibbs, 1974) also point out that these estimates are not efficient, however, and that statistical inference when the endogenous residuals are autocorrelated is extremely hazardous.

5 Dynamic autocorrelated model

In the first part of this paper, I developed a model for a dynamic system which, in addition to other features, was not a function of the initial state of the system. In effect, this assumption represents the claim that the present state of the system is entirely a function of the structure of the system: repeated reenactment of the dynamic sequence for a given structure (i.e., a given coefficient matrix and given disturbance variance–covariance matrix) would result in an identical expected state of the system for any time t, whatever the initial state or whatever the time since the initiation of the process. For many social processes, of course, this will not be a desirable assumption, in particular for very small-scale processes such as those generated in the laboratory. However, for many large-scale social processes, I believe that explaining the present state of the system in terms of the initial state is attributing important consequences to essentially fortuitous events.

A term often used to distinguish these two kinds of dynamic processes is "ergodic" (Feller, 1966, p. 208). The ergodic dynamic process "forgets" the past, and the nonergodic process is a function of its initial state. As we have seen in the previous section of this paper, the static regression model, with and without the autocorrelation of the endogenous disturbance term, is nonergodic. Thus, these models are inappropriate when we assume that the data have been generated by an ergodic dynamic process.

When we relax our assumptions about the behavior of the exogenous variables, however, we will find that an ergodic process results. We have assumed previously that $X_t = X_{t-1} + V_t$, where V_t is a matrix of serially uncorrelated random variables. Let us, instead, specify that

$$X_t = \theta X_{t-1} + V_t \tag{14}$$

where V_t retains the properties of serial uncorrelation and randomness, and θ is a diagonal coefficient matrix of autoregression coefficients. Furthermore, let $y_t = BX_{t-1} + u_t$ where $u_t = \phi u_{t-1} + u_t^*$. Then, by the results of the previous section:

$$y_t = BX_{t-1} + \phi y_{t-1} - \phi BX_{t-2} + u_t^* \tag{15}$$

putting equations (14) and (15) directly into the form of equation (1) results in a trivial singularity. Instead, lagging equation (14) one time period, $X_{t-1} = \theta X_{t-2} + V_{t-1}^*$, and solving for X_{t-2}, we get $X_{t-2} = \theta^{-1}(X_{t-1} - V_{t-1}^*)$; let $w_t^* = \phi B\theta^{-1}V_{t-1}^* + u_t^*$ and insert these results into equation (15) to obtain:

$$y_t = \phi y_{t-1} + (B - \phi B\theta^{-1})X_{t-1} + w_t^* \tag{16}$$

Now we may put equations (14) and (16) into the form of equation (1):

$$\begin{bmatrix} y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \phi & (B - \phi B\theta^{-1}) \\ 0 & \theta \end{bmatrix} \begin{bmatrix} y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} w_t^* \\ V_t^* \end{bmatrix} \tag{17}$$

The characteristic roots of the coefficient matrix of this equation are equal to ϕ and the diagonal elements of θ. Since negative values of ϕ and θ would be difficult to interpret, ϕ and the elements of θ must be greater than zero, and, if the system is to be stable, they must also be less than one.

Given that this is so, we may calculate the variance–covariance matrix of the variables at time t using equation (5). Partitioning Σ in equation (5) thus,

$$\Sigma = \begin{bmatrix} \Sigma_{yy} \Sigma_{yx} \\ \Sigma_{xy} \Sigma_{xx} \end{bmatrix} \tag{18}$$

and

$$E(\Xi) = \begin{bmatrix} E(w_t^* w_t^{*\prime}) \mathrm{E}(w_t^* V_t^{*\prime}) \\ E(V_t^* w_t^{*\prime}) \mathrm{E}(V_t^* V_t^{*\prime}) \end{bmatrix}$$

but $E(V_t^* w_t^{*\prime}) = E[V_t^*(V_{t-1}^{*\prime}\theta^{-1}B'\phi + u_t^*)] = 0$, since, by hypothesis, the residuals in V_t^* are serially uncorrelated and are uncorrelated with the residuals in u_t^*. Therefore,

$$\mathrm{E}(\Xi) = \begin{bmatrix} E(w_t^* w_t^{*\prime}) & 0 \\ 0 & E(V_t^* V_t^{*\prime}) \end{bmatrix} = \begin{bmatrix} \Psi & 0 \\ 0 & \Omega \end{bmatrix} \tag{19}$$

Substituting equations (18) and (19) and the coefficient matrix in equation (17) into equation (5), and solving for Σ_{yx}, we get $\Sigma_{yx} - \phi\Sigma_{yx}\theta - (B - \phi B\theta^{-1})\Sigma_{xx}\theta = 0$. Postmultiplying by Σ_{xx}^{-1} and using $\Sigma_{xx}^{-1}\Sigma_{xx} = I$,

$$\Sigma_{yx}\Sigma_{xx}^{-1} - \phi\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xx}\theta\Sigma_{xx}^{-1} - (B - \phi B\theta^{-1}) \\ \times \Sigma_{xx}\theta\Sigma_{xx}^{-1} = 0$$

but, since $\hat{\Gamma} = \Sigma_{yx}\Sigma_{xx}^{-1}$ and letting $\rho = \Sigma_{xx}\theta\Sigma_{xx}^{-1}$ we get $\hat{\Gamma} - \phi\hat{\Gamma}\rho = (B - \phi B\theta^{-1})$, and since ϕ is a scalar, $\hat{\Gamma}(I-\phi\rho) = B(I-\phi\theta^{-1})\rho$; postmultiplying by $(I-\phi\rho)^{-1}$, $\hat{\Gamma} = B(I-\phi\theta^{-1})\rho(I-\theta^{-1})\rho(I-\phi\rho)^{-1}$. If we let the q autocorrelation coefficients in θ be all equal to, say, μ, then $\rho = \Sigma_{xx}\theta\Sigma_{xx}^{-1} = \mu\Sigma_{xx}\Sigma_{xx}^{-1} = \mu I$ and $\hat{\Gamma} = [(\mu-\phi)/(1-\mu\phi)]B$.

The parameters of our static multiple regression model, therefore, are proportional to the parameters of the analogous autocorrelated dynamic process, in which the constants of proportion are fairly complicated functions of the autocorrelation coefficients and the variances and covariances of the independent variables in the general case, and are somewhat simple functions of the autocorrelation coefficients in the case where those of the exogenous variables are all equal. Note also that, if the autocorrelation coefficients are all about equal in value—a perfectly reasonable possibility—then the regression coefficients calculated in the multiple regression model will be severely attenuated versions of the analogous dynamic process.

6 A First-order dynamic model

Stability of the variables in the model may also be produced by an explicit first-order autoregressive dynamic process. In the previous section, an essentially zero-order process became a first-order process as a result of the autocorrelation. Here, we shall consider a model which is hypothesized directly to be a first-order process without autocorrelation. In the first-order dynamic process, we claim that the *rate* of the dependent variable is a function of the dependent variable at the previous point in time as well as of the independent variables and the residual variables. For example, $\Delta y_t = \alpha y_{t-1} + bx_{t-1} + u_t$ and $\Delta x_t = cx_{t-1} + \nu_t$, but $\Delta y_t = y_t - y_{t-1}$ and $\Delta x_t = x_t - x_{t-1}$ and, therefore, $y_t = (1+\alpha)y_{t-1} + bx_{t-1} + u_t$ and $x_t = (1+c)x_{t-1} + \nu_t$ is a first-order autoregressive dynamic structural model incorporating one dependent and one independent variable. This is the model most commonly specified in multiple-wave studies (e.g., Heise, 1970; Jöreskog and Sörbom, 1975), where $1+\alpha$ and $1+c$ are interpreted as "stability coefficients." Figure 16.1 will make this more apparent for those more familiar with path diagrams. It diagrammatically represents the aforementioned first-order autoregressive equations.

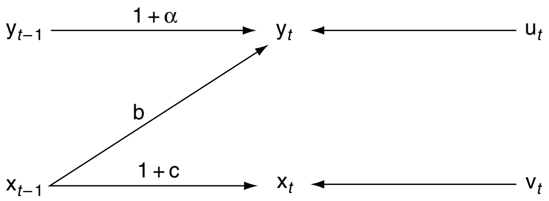


Figure 16.1 Path diagram of a first-order auto regressive dynamic process

These equations may be generalized to any number of endogenous and exogenous variables:

$$\Delta Y_t = AY_{t-1} + BX_{t-1} + U_t \tag{20}$$

$$\Delta X_t = CX_{t-1} + V_t \tag{21}$$

where A contains coefficients of the relationships among the endogenous variables off the diagonal and the autoregressive coefficients on the diagonal; B contains the coefficients of the relationships of the endogenous to the exogenous variables; C is a diagonal matrix of autoregressive coefficients of the exogenous variables; Y_t is the matrix of endogenous variables; X_t is the matrix of exogenous variables; and U_t and V_t are matrices of residuals with the usual properties.

In the analogous static or cross-sectional structural model let $\Pi\mathrm{Y} = \Gamma X + Z$, where Π is the matrix of regression coefficients of the endogenous variables in Y on the other endogenous variables (with the diagonal normed to 1); Γ is the matrix of regression coefficients of the endogenous variables in Y on the exogenous variables in X; and Z is the matrix of residual variables. Partitioning Σ, the variance–covariance matrix of the variables, into endogenous variables first and exogenous variables second, we have

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} = \begin{bmatrix} E(YY') & \mathrm{E}(YX') \\ E(XY') & \mathrm{E}(XX') \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\Pi}^{-1}\hat{\Gamma}\hat{\Sigma}_{xx}\hat{\Gamma}'\hat{\Pi}'^{-1} + \hat{\Pi}^{-1}\hat{\zeta}\hat{\Pi}'^{-1} & \hat{\Pi}^{-1}\hat{\Gamma}\hat{\Sigma}_{xx} \\ \hat{\Sigma}_{xx}\hat{\Gamma}'\hat{\Pi}'^{-1} & \hat{\Sigma}_{xx} \end{bmatrix} \tag{22}$$

where $\zeta = E(ZZ')$. Since $\hat{\Sigma}_{xx}$ is ordinarily unconstrained, then $\hat{\Sigma}_{xx} = \Sigma_{xx}$.

Rewriting equations (20) and (21):

$$Y_t = (I + A)Y_{t-1} + BX_{t-1} + U_t \text{ and}$$

$$X_t = (I + C)X_{t-1} + V_t$$

therefore,

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} I+A & B \\ 0 & I+C \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} U_t \\ V_t \end{bmatrix} \tag{23}$$

The characteristic roots of the coefficient matrix in equation (23) are equal to the characteristic roots of $I + A$ and $I + C$, respectively, since

$$\begin{vmatrix} I+A & B \\ 0 & I+C \end{vmatrix} = |I + A| \cdot |I + C|$$

The roots of $I + C$ will be less than one in modulus if the nonzero elements of C are less than zero. The nonzero elements of C, which is a diagonal matrix, will therefore be negative. The roots of $I + A$ depend on the configuration of A. If A represents the coefficients of an exactly recursive set of equations, for instance, the elements of A may be arranged so that the upper right portion of the matrix will be all zeros. The roots will then be the diagonal elements of $I + A$, and, necessarily, the diagonal elements of A will be negative if the roots are to be less than one in modulus.

Let

$$\Psi = E(U_t U_t'), \Omega = E(V_t V_t'), \text{ and } \Xi = \begin{bmatrix} \Psi & 0 \\ 0 & \Omega \end{bmatrix}$$

Substituting this equation, the partitioned coefficient matrix in equation (23), and Σ, partitioned as in equation (22), into equation (5), and solving for Σ_{yx}, we get $\Sigma_{yx} - (I + A)\Sigma_{yx}(I + C) - B\Sigma_{xx}(I + C) = 0$. Postmultiplying by Σ_{xx}^{-1} and using $\Sigma_{xx}^{-1}\Sigma_{xx} = I$:

$$\Sigma_{yx}\Sigma_{xx}^{-1} - (I + A)\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xx}(I + C)\Sigma_{xx}^{-1}$$
$$- B\Sigma_{xx}(I + C)\Sigma_{xx}^{-1} = 0 \tag{24}$$

Letting $\rho = \Sigma_{xx} C \Sigma_{xx}^{-1}$ then $I + \rho = \Sigma_{xx}(I + C)\Sigma_{xx}^{-1}$. From equation (22) we have $\Sigma_{yx}\Sigma_{xx}^{-1} = \hat{\Pi}^{-1}\hat{\Gamma}$, where $\hat{\Pi}$ and $\hat{\Gamma}$ are the parameter matrices of the nondynamic structural model. Substituting these results into equation (24):

$$\hat{\Pi}^{-1}\hat{\Gamma} - (I + A)\hat{\Pi}^{-1}\hat{\Gamma}(I + \rho) - B(I + \rho) = 0 \tag{25}$$

If the nondynamic structural model in equation (22) is specified such that it is exactly identified and that ζ is unconstrained, then equation (25) contains enough equations to determine uniquely the parameters in $\hat{\Pi}$ and $\hat{\Gamma}$ in terms of the parameters in A, B, and ρ. These expressions will necessarily be very complicated. They may be considerably simplified if we let the nonzero coefficients in C be equal to, say, μ. Then, $I + \rho = (1 + \mu)I$, and equation (25) becomes $[1/(1 + \mu)]\hat{\Pi}^{-1}\hat{\Gamma} - (I + A)\hat{\Pi}^{-1}\hat{\Gamma} = B$ and $\hat{\Pi}^{-1}\hat{\Gamma} = [\mu/(1 + \mu)I + A]^{-1}B$. Now, let P be a matrix of the off-diagonal elements of A with zeros on the diagonal, and let the coefficients on the diagonal of A be equal to, say, α; then $P + \alpha I = A$ and $\hat{\Pi}^{-1}\hat{\Gamma} = [\mu/(1 + \mu)I + \alpha I + P]^{-1}B = [(\alpha + \mu/(1 + \mu))I + P]^{-1}B$. Let $\tau = \alpha + \mu/(1 + \mu)$, then

$$\hat{\Pi}^{-1}\hat{\Gamma} = [I + (1/\tau)P]^{-1}(1/\tau)B \tag{26}$$

Assuming again that the specified structural equation model in equation (22) is identified exactly and that ζ is unconstrained, then we may deduce from equation (26) that

$$\hat{\Pi} = I + (1/\tau)P \tag{27}$$

$$\hat{\Gamma} = (1/\tau)B \tag{28}$$

since the multiplication of the equations in equation (26) would result simply in m equations of the following form, where m is the number of parameters in $\hat{\Pi}$ and $\hat{\Gamma}$ that are being estimated: $\hat{\pi}_{ij} = p_{ij}/\tau$ and $\hat{\gamma}_{ij} = b_{ij}/\tau$, where $\hat{\pi}_{ij}, p_{ij}, \hat{\gamma}_{ij}$, and b_{ij} are the ij-th elements of $\hat{\Pi}, P, \hat{\Gamma}$, and B, respectively. If the model is overidentified or if the elements of ζ must be constrained to identify the model, the results are more complex.

We see, then, that the estimated parameters of our static structural equation model are, in this case, simple proportions of the corresponding dynamic parameters, when in fact the underlying model that generated the data is dynamic. The constant of proportionality is a function of the autoregression coefficients of the dynamic model. Without making the argument too unrealistic, we may assume that the autoregression coefficients are all about equal to each other. Let that magnitude be represented by δ. Then

$$1/\tau = (1+\delta)/[\delta(2+\delta)] \qquad (29)$$

Because of the manner in which the dynamic model was specified in equations (20) and (21), δ must be less than 0 and greater than −1. A quick glance at equation (29) reveals that, as δ approaches 0, i.e., as the autoregression coefficients approach 0 together, the constant of proportionality, $1/\tau$, approaches 0, and the estimated cross-sectional coefficients become very attenuated versions of the corresponding dynamic coefficients.

7 Conclusion

Whether we like it or not, sociological variables are imbedded in time; and whether methodologically convenient or not, we must ensure that our models properly reflect that fact. Experimental methodology, when appropriate, includes the investigator's ability to manipulate the independent variables and to control to some extent the nature of the endogenous residual variation, thus making reasonable the assumptions of fixed independent variables and serially uncorrelated residual variation, and, therefore, making reasonable the specification of static fixed effects models. Such manipulation and control is not possible in nonexperimental methodology, however, rendering unreasonable the assumption of fixed exogenous variables and severely narrowing the instances in which the static model is appropriate. When considering a static model for time-dependent sociological variables, we must be prepared to justify claims regarding the behavior of these variables *across time*, and the success of cross-sectional analysis of sociological variables will depend on our success in defending such claims.

The analysis of static and dynamic models presented in this paper point toward an inevitable conclusion: that efficient unbiased estimates of structural coefficients using static models are possible only when we are prepared to assume that the underlying process that has generated the data is nonergodic, i.e., that the process is a function of the initial state of the system; and that change in the exogenous variables at any point in time is the same. If, however, we wish to assume that the influence of the past vanishes in proportion to its distance from the present, that the present state of the system is entirely a function of the structure of the system and is not a function of the initial state of the system, then the system is ergodic, i.e., it has "forgotten" the past. Such a system, we have seen, cannot be successfully investigated by the use of static regression models. Attempts to apply static models to ergodic dynamic processes have resulted in estimates of static parameters that are attenuated or inflated versions of the parameters of the underlying dynamic process.

Because most social variables are inherently embedded in time and, therefore, are generated by dynamic processes, we must pay strict attention to the nature of these processes. And, if we are to continue to apply static structural models to these processes, we must be prepared to argue that they are nonergodic.

References

Balestra, P. and Nerlove, M. (1966). Pooling cross-section and time-series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica*, 34: 585–612.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, vol. 2. New York: Wiley.

Goldberger, A. S. (1964). *Econometric Theory*. New York: Wiley.

Heise, D. R. (1969). Separating reliability and stability in test–retest correlation. *American Sociological Review*, 34: 93–101.

Heise, D. R. (1970). Causal inference from panel data. In E. F. Borgotta and G. W. Bohrnstedt (eds), *Sociological Methodology*, pp. 3–27. San Francisco: Jossey-Bass.

Heise, D. R. (1975). *Causal Analysis*. New York: Wiley.

Hibbs, D. A., Jr. (1974). Problems of statistical estimation and causal inference in time-series regression models. In H. L. Costner (ed.), *Sociological Methodology*, Chapter 9, pp. 252–308. San Francisco: Jossey-Bass.

Jöreskog, K. G. and Sörbom, D. (1975). Statistical models and methods for analysis of longitudinal data. *Research Reports* 75: 1, University of Uppsala, Sweden.

Kuh, E. (1959). The validity of cross-sectionally estimated behavior equations in time-series applications. *Econometrica*, 27: 197–214.

Miller, K. S. (1968). *Linear Difference Equations*. New York: W. A. Benjamin.

Schoenberg, R. J. (1976). The estimation of dynamic parameters in cross-sectional data. Photocopy.

Theil, H. and Goldberger, A. S. (1961). On pure and mixed statistical estimation in economics. *International Economic Review*, 2: 65–68.

Wiley, D. E. and Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, 35: 112–117.

Young, K. A. (1972). A synthesis of time-series and cross-section analyses: Demand for air transportation service. *Journal of the American Statistical Association*, 67: 500–566.

This page intentionally left blank

Chapter 17

Causal analysis with nonexperimental panel data

David F. Greenberg

1 Introduction

This chapter surveys methods for conducting causal analysis with non-experimental panel data (data containing observations for multiple cases at multiple, evenly-spaced times). It begins by describing the criteria used in empirical social science research for establishing that one variable is the cause of another. It then takes up, in turn, the analysis of models for qualitative outcomes, and models for quantitative outcomes. It concludes with a discussion of causal inference from independent cross-sections. To keep the discussion manageable, the chapter ignores problems arising from measurement error and missing data.

2 Causal analysis with panel data

For purposes of this chapter, the variable X is considered to be a cause of variable Y when three conditions are met: there exists some type of association between the two variables (e.g., a non-vanishing correlation or partial correlation), X precedes Y in time, and there is no other explanation for the association. This last condition means, in particular, that the association is not spurious.

An assessment of the first criterion can be carried out with cross-sectional data (data for a single time), but that is not true of the second condition. As Paul Lazarsfeld and Marjorie Fiske (1938) pointed out long ago, in a cross-sectional research design, all variables are measured at the same time. Consequently, there is no way of knowing that a putative cause came before its supposed effect. With data measured at two or more points in time, the time-ordering of the observations is not in question.

The gains to be had from panel data, Paul Lazarsfeld (1972) noted, are especially great when theoretical reasons can be found for thinking that the causal influences between two variables can occur in either direction. In that circumstance, it may be unclear whether a correlation between X and Y exists because X causes Y, Y causes X, or other variables cause both X and Y. While reciprocal causal effects can be distinguished with cross-sectional data, the statistical procedures for doing this (e.g., two-stage least-squares) require strong, untestable assumptions about the effects of exogenous variables on the endogenous variables. Sometimes these assumptions are implausible, making their use a dubious proposition. With panel data, it would seem, this difficulty should not arise.

Over the past half-century, methodologists have developed procedures for carrying out causal analyses with panel data. This chapter surveys the major methods now available for doing this, and points to some of the issues and pitfalls that arise in their use.

3 Qualitative outcomes

Lazarsfeld's (1972) method for determining the predominant direction of causal influence between two variables may have been the first statistical method developed for the particular purpose of analyzing panel data. We illustrate the method with his example: voter opinion in the 1940 presidential campaign. Surveyers asked a panel of voters in June, and again in August, to specify the party for which they intended to vote in the November election, and to express an opinion of Republican candidate Wendell Wilkie. As expected, Republican voters were more likely than Democrats to think well of Wilkie, but from this alone, one would have no way of knowing whether this was because Republicans were more likely to think favorably of their party's nominee, whoever it was; or because voters who liked Wilkie were more likely to vote for his party. Lazarsfeld's method was designed to find out which influence was stronger: the effect of party preference on respondents' evaluations of the candidate, or the effect of their assessments of the candidate on party preference.

Results of the survey are displayed in Table 17.1, a 16-fold table constructed by cross-tabulating preferences and opinions at the two times. Comparing the row totals, which measure intentions and preferences at the first wave, with the column totals, which represent intentions two months later, we see that some shift has taken place. In August there were fewer respondents expressing incongruent responses (those holding favorable opinions of Wilkie, but planning to vote Democratic, and those holding unfavorable opinions of Wilkie but planning to vote Republican) than in June.

All the entries off the main diagonal (which runs from upper left to lower right) represent respondents whose views of Wilkie, or whose voting intentions (or both), changed between wave 1 and wave 2. The entries on the minor diagonal (which runs from lower left to upper

Table 17.1 Intention to vote and opinion of Wilkie

First wave	*Second wave*				
	Democrat against Wilkie	*Democrat for Wilkie*	*Republican against Wilkie*	*Republican for Wilkie*	*Total*
Democrat against Wilkie	68 (f_{11})	$\underline{2}^{**}$ (f_{12})	$\underline{1}^{**}$ (f_{13})	1 (f_{14})	72 (n_1)
Democrat for Wilkie	$\underline{11}^{*}$ (f_{21})	12 (f_{22})	0 (f_{23})	$\underline{1}^{*}$ (f_{24})	24 (n_2)
Republican against Wilkie	$\underline{1}^{*}$ (f_{31})	0 (f_{32})	23 (f_{33})	$\underline{11}^{*}$ (f_{34})	35* (n_3)
Republican for Wilkie	2 (f_{41})	$\underline{1}^{**}$ (f_{42})	$\underline{3}^{**}$ (f_{43})	129 (f_{44})	135 (n_4)
Total	**82**	**15**	**27**	**142**	**266**

Note: the meaning of the asterisks and underlinings is explained in the text.

right) represent respondents who changed both their party preference and their opinion of Wilkie. These cases do not furnish information about the direction of causal influence because they represent either respondents who did not change on either variable, or respondents who changed on both. Consequently, they provide no information as to which changed first, party preference or opinion of the candidate. Lazarsfeld ignores these cases, and concentrates on the eight cells with underlined frequencies. These cases all involve change on one variable but not the other.

First, consider the underscored cases in the second and third rows of the table. These represent cases of incongruous response at the first wave. They are Democrats who liked the Republican candidate (row 2) and Republicans who disliked the Republican candidate (row 3). There are two ways incongruent responders could become congruent responders. Democrats could maintain their party preference, and come to dislike the Republican candidate (a response chosen by 11) , or they could maintain their opinion of the candidate but change their party affiliation (a response of 1 subject). Similarly, incongruent Republicans could become congruent responders by maintaining their party preference but coming to like Wilkie, or by maintaining their dislike for Wilkie but changing their party preference. All these movers from incongruous to congruous responses are marked by a single asterisk.

The first and fourth row of the table shows responses that are initially congruent. They are Democrats who disliked Wilkie, and Republicans who liked him. A double asterisk marks those who moved from congruence to incongruence by abandoning their party or revising their views of the candidate.

Lazarsfeld argues that if the influence of party preference on candidate assessment is stronger than the reverse influence, more of the respondents will maintain their party preference but alter their views of the candidate than vice versa. If the influence of candidate assessment is stronger, more of the respondents will maintain their view of the candidate but alter their party preference. A significance test can be carried out to determine whether the difference in change patterns is greater than expected on the basis of chance, given the row and column totals.

This method is limited to pairs of binary variables. Variables with more than two categories could be accommodated only by dichotomizing, a procedure that loses information. In addition, the procedure becomes cumbersome when additional variables need to be taken into account. Consequently, energies have been invested primarily in methods for handling continuous variables. These methods will be reviewed in Section 4 of this chapter.

Nevertheless, a few researchers have found distinct advantages in working with dichotomous variables. Ronald Kessler (1977), for example, observed that linear models for continuous variables treat increases and decreases in a variable symmetrically. However, it is possible that party preference has a different effect on prospective voters who dislike a candidate than it does on those who like that candidate. Linear panel models (models for continuous variables), however, treat those effects as symmetrical. The analysis of cross-tabulations is well-suited to the study of asymmetrical influences.

The additional information that can be extracted from the 16-fold table through strategic comparisons of cell frequencies can be seen by considering Kessler's analysis of a two-wave survey of high school students surveyed in the Fall of 1971, and again in the Spring of 1972. Table 17.2 summarizes the responses to a "yes–no" question about marijuana use, and a scale of depression that has been dichotomized at the median.

Table 17.2 Marijuana use and depression

First wave	*Second wave*				
	User depressed	*User undepressed*	*Nonuser depressed*	*Nonuser undepressed*	*Total*
User depressed	225 (f_{11})	110 (f_{12})	78 (f_{13})	45 (f_{14})	488 (n_1)
User undepressed	78 (f_{21})	243 (f_{22})	15 (f_{23})	77 (f_{24})	413 (n_2)
Nonuser depressed	159 (f_{31})	58 (f_{32})	1019 (f_{33})	482 (f_{34})	1718 (n_3)
Nonuser undepressed	44 (f_{41})	135 (f_{42})	401 (f_{43})	1586 (f_{44})	2166 (n_4)
Total	**536**	**546**	**1513**	**2190**	**4785**

We consider the following four comparisons:

depressed users becoming undepressed versus depressed nonusers becoming undepressed

$$(f_{12}+f_{14})/n_1 = (110+45)/488 = .318$$

versus

$$(f_{32}+f_{34})/n_3 = (58+482)/1718 = .314$$

$$\Delta = .318 - .314 = .004$$

undepressed users becoming depressed versus undepressed nonusers becoming depressed

$$(f_{21}+f_{23})/n_2 = (78+15)/413 = .225$$

versus

$$(f_{41}+f_{43})/n_4 = (44+401)/2166 = .205$$

$$\Delta = .225 - .205 = .020$$

depressed users stopping use versus undepressed users stopping use

$$(f_{13}+f_{143})/n_1 = (78+45)/488 = .252$$

versus

$$(f_{23}+f_{24})/n_2 = (15+77)/413 = .223$$

$$\Delta = .252 - .223 = .029$$

depressed nonusers initiating use versus undepressed nonusers initiating use

$$(f_{31}+f_{32})/n_3 = (159+58)/1718 = .126$$

versus

$$(f_{41}+f_{42})/n_4 = (44+135)/2166 = .083$$

$$\Delta = .126 - .083 = .043$$

The differences are all quite small, indicating that the initial mental state and drug use status have at most weak effects on drug use and depression at a later time. Ignoring the small size of the effects, it appears from the first of the comparisons that among those who were depressed, marijuana users were a bit more likely than nonusers to lose their depression. The second comparison shows that among those who were not depressed, marijuana users were more likely than nonusers to become depressed. The third comparison shows depressed users to be more likely than users who are not depressed to stop using. Lastly, depression raises the likelihood that a nonuser will initiate marijuana use.

To determine the role of third variables in maintaining an initial pattern, for each row in Table 17.2 we construct a 2×2 table, always placing the frequency of cases in which no change in either variable is experienced in the upper left cell, the frequency of cases in which both variables change in the lower right cell, and the frequency of cases for which just one variable changes in the off-diagonal cells. The tendency of the time-1 cross-sectional distribution to be preserved at time-2, above and beyond the level required by the row and column distributions, can be assessed by the Q statistic, defined by the equation

$$Q = \frac{ad - bc}{ad + bc} \qquad (1)$$

For the four rows in Table 17.2, the Q statistic is .083, −.245, −.129 and .126. The significance of each of the cross-lagged effects and these maintenance effects can be tested by fitting a Goodman log-linear model to the data. Todd Miller and Brian Flay (1996) have developed an alternative strategy for analyzing causal relationships from categorical data, specifically designed to examine stage-like phenomena (where one type of event – such as smoking tobacco cigarettes – is a stepping-stone to another type of event – such as smoking marijuana).

The methods just summarized fail to take into account the length of time during which a variable changes. James Coleman's (1964) method for analyzing panel data does just that. Assuming that change can take place at any time between the first and second wave, we write linear differential equations representing the flow of cases from one state to another. If, for example, there are four states defined by the combination of two binary variables (such as the use or nonuse of marijuana and the presence or absence of depression), the equation representing flows of cases into state 1 would be

$$\frac{dn_1}{dt} = r_{12}n_2 + r_{13}n_3 + r_{14}n_4 - (r_{21} + r_{31} + r_{41})n_1 \qquad (2)$$

Here r_{ij} represents the instantaneous rate at which case are moving from state j into state i. The negative signs for the coefficient in parentheses indicates that they represent transitions out of state 1 into the other three states. Similar equations can be written for n_2, n_3 and n_4. With two waves of data, the 16 coefficients can be estimated from the 16 observed frequencies. The computations are too complicated to cover here, but Coleman (1964) shows how approximate solutions can be obtained.

4 Quantitative (interval-level) outcomes

4.1 Cross-lagged panel correlations

The earliest method for carrying out a causal analysis with interval-level panel data was inspired by Lazarsfeld's procedures for analyzing dichotomous data. The method, known by its acronym CLPC (cross-lagged panel correlations), tries to answer the question, "Which is the more important influence, X on Y, or Y on X?", when measurements on these variables are available at just two points in time. As initially formulated, the method compares the correlation between X measured at time 1 and Y measured at time 2 with the correlation between Y measured at time 1 and X measured at time Z. Using subscripts to identify the wave, a comparison is made between $r_{X_1Y_2}$ and $r_{Y_1X_2}$. The correlation that is largest in magnitude identifies the stronger influence (Campbell and Stanley, 1963). Originators of the method boasted that it could yield conclusions about causality on the basis of nonexperimental data that were virtually as trustworthy as those achieved with an experimental design involving random assignment of subjects to different conditions (a procedure that eliminates spuriousness as a possible explanation for a relationship between two variables).

Relieved of concerns about spuriousness that had up to then vexed researchers working with nonexperimental data, flocks of researchers adopted the new method. Soon, however, methodological critiques established that the CLPC method worked only under special conditions that did not always hold. These assumptions were:

1. that the relationships being studied be stationary (cross-sectional correlations not changing over time) and
2. that the autocorrelations of each variable be equal. This means that $r_{X_1X_2}$ had to equal $r_{Y_1Y_2}$.

Some researchers modified the procedure so that the CLPC criterion involved a comparison between the partial correlations $r_{X_1Y_2Y_1}$ and $r_{Y_1Y_2X_1}$ rather than the zero-order correlations (Pelz and Andrews, 1964). However, the development of more powerful statistical methods led most researchers to abandon the CLPC strategy as a method for analyzing causal influences.

David Kenny (1975; 1979, pp. 235–39) resurrected the comparison of cross-lagged panel correlations as a test for spuriousness. If the two cross-lagged correlations are not significantly different from one another, he noted, the relationship between *X* and *Y* is spurious. Under conditions of stationarity this is a valid test for spuriousness. As we will see, structural equation modeling leads to an additional test of spuriousness that is valid whether or not stationarity holds. Consequently, the CLPC is of limited value as a test for spuriousness.

Faced with the limitations of the CLPC method, most researchers turned to other methods for studying causal relationships with panel data. These methods include structural equation modeling (SEM), the pooling of cross-sections, and latent growth curve modeling.

4.2 Causal analysis in structural equation modeling

SEM begins by postulating a linear causal model based on theoretical considerations. For a two-wave panel data set, one begins by writing a linear structural equation for a continuous, interval-level dependent variable *y* at time 2. If *x* is an independent variable, the most general equation will utilize *y* at time 1, and *x* at times 1 and 2, as predictors of *y* at time 2, with random error term *e*:

$$y_{i2} = a + b_1 y_{i1} + b_2 x_{i1} + b_3 x_{i2} + e_i \qquad (10)$$

Where it is not needed for clarity, the first index, designating the case, will be suppressed. Additional predictors can be added easily; we will concentrate on the two-variable case to avoid complicating the equations unnecessarily. If there are more waves, one can write additional equations for them, in each case using the lagged endogenous variable and the contemporaneous and lagged exogenous variables as predictors.

Ideally, the time between two waves should correspond roughly to the length of time for significant change in the dependent variable to occur. Someone who tried to determine the influence of smoking cigarettes on lung cancer would find none if analyzing observations taken a day apart, because it typically takes years before smoking leads to cancer.

The interpretation of the coefficients in equation (10) can be clarified by subtracting the lagged value of y_1 from both sides of the equation. Denoting the difference operator by Δ, we have

$$\Delta y_i = y_{i2} - y_{i1} = a + (b_1 - 1) y_{i1} + b_2 x_{i1} + b_3 x_{i2} + e_i \qquad (11)$$

The left-hand side in this equation represents change in *y*. Comparing the coefficients in this equation with those in equation (10), we see that all the coefficients are unchanged, except for the coefficient of the time 1 (lagged) endogenous variable. This coefficient is the original coefficient minus 1. It follows that we can go back and forth between an equation for the level of the dependent variable in the second wave, and an equation for the change in that variable between the first and second waves, with the simplest algebraic maneuver.

If we wish to interpret these equations as expressing a causal relationship, then a represents a constant change for all cases between the first and second wave. The coefficient $b_1 - 1$ represents the effect of the value of y at time 1 on change in y. One might wonder whether it is legitimate to think of y at time 1 as causing y at time 2. In many circumstances it is quite plausible to do so. Someone who has been victimized by a crime may respond by retaliating against the perpetrators, so that crime leads to crime. Actions at a given time may result in positive or negative reinforcement that enhances or reduces the likelihood of that activity in the future. A student's mastery of mathematics skills at a particular time may provide the foundation for further learning of more advanced mathematics topics, and so may prove to be a cause of mathematical competence at a later age. Children who have many friends may, by interacting with them, acquire social skills that enable them to widen their friendship networks. In circumstances like this, it is meaningful to consider a variable to be influencing its own level at a later time.

The coefficient b_2 is the effect of x at time 1 on change in y. For example, the level of crime in a community could stimulate growth in the police force. The number of hours children spend doing homework or watching television could influence the rate at which their grades improve. The level of unemployment in a community may influence the rate at which prices rise.

If taken at face value, the coefficient b_3 expresses the effect of x at time 2 on the change y undergoes between times 1 and 2. Yet, the principle of causality insists that an event at a particular time can only influence events at future times. Notwithstanding this seeming difficulty, many researchers use this type of specification to represent the contemporaneous influence of one variable on change in another. Because time is often measured crudely, this term can be taken to represent the effect of influences that act quickly on a timescale set by the interval between waves. An alternative way of understanding this coefficient is to realize that the equality

$$x_2 = x_1 + \Delta x \tag{12}$$

allows us to reparametrize equation (10) as expressing the lagged effect of x on y and the effect of change of x on change in y.

If the coefficients in equation (10) remain constant for a long period, it is possible to project the long-term behavior of the system. It is easy to show that y will asymptotically approach an equilibrium value

$$y_{EQ} = \frac{a + (b_2 + b_3)x}{1 - b_1} \tag{13}$$

provided that the coefficient b_1 is less than 1. We see that the effect of a change in x will be magnified by the factor $1/(1 - b_1)$. This can be large. Consequently, the coefficients b_2 and b_3, which represent the short-term effects of a change of x on y, give an incomplete view of the full, long-term consequences of change.

In some research problems, there may be persuasive theoretical reasons for thinking that one variable could influence another with a lag and also contemporaneously, and that there could, in addition, be a "change causes change" contribution. In this case, the researcher may be tempted to introduce all three terms (x_1, x_2 and Δx) as predictors. Yet, because of the perfect linear dependence of these three variables, it is impossible to do this, at least if all influences are thought to be linear, and no additional information about their effects is available. Trying to do so would lead to perfect multicollinearity.

At times, this circumstance leads to genuine ambiguity as to the nature of the causal process at work. At other times, however, some possibilities will seem more plausible than others. Consider the relationship between indices measuring time spent socializing and involvement in theft and vandalism, in the first two waves of the Youth in Transition study of a nationally representative sample of high school students (Rosenberg and Rosenberg, 1978). The means, standard deviations, and correlations among these variables, measured a year apart, are shown in Panel A of Table 17.3. As expected, each time-2 variable is positively correlated with its own time-1 value. The cross-sectional correlations between social life and theft-vandalism are positive, as are the cross-lagged coefficients. The mean social life score rose slightly over the course of a year, while involvement in theft and vandalism dropped slightly.

Let us use these data to assess the influence of social life on theft. If we regress *THEFT* at time 2 on *THEFT* at time 1, and social life at times 1 and 2, we obtain the results shown in Panel B of Table 17.3. As expected, the effect of *THEFT* at time 1 on *THEFT* at time 2 is positive. The influence of initial levels of theft on change in theft is measured by .481 − 1 = −.519, which is substantially negative. We thus conclude that involvement in *THEFT* tended to decline more for those subjects with initially high levels of involvement than for the group as a whole, and it tended to decline less for those with initially low levels of involvement than for the group as a whole. The lagged and contemporaneous effects of social life are both positive but quite

Table 17.3 The relationship between undesirable life events and self-esteem

Panel A Correlations among the variables at two points in time[a]

Variable	*Variable*					
	SL1	*THEFT1*	*SL2*	*THEFT2*	*Mean*	*SD*
SL1	1				333.368	93.999
THEFT1	.22	1			153.097	54.509
SL2	.51	.16	1		371.566	92.516
THEFT2	.18	.47	.17	1	137.193	48.945

[a] $n = 1412$. *SL* = social life, *THEFT* = theft and vandalism. *Source*: Bynner, O'Malley and Bachman (1981).

Panel B Regression of $THEFT_2$ on $THEFT_1$, SL_1 and SL_2

Variables	*b*	*s.e.*	*beta*	*t*	*sig*
constant	49.969	5.803		8.611	.000
THEFT1	.418	.022	.448	18.693	.000
SL1	.023	.015	.042	1.533	.125
SL2	.042	.015	.077	2.820	.005

$R^2 = .231$

small. Only the contemporaneous effect is statistically significant at the .05 level.[1]

Taken at face value, these results suggest that the effect of undesirable life events on theft tends to be short-lived. Yet one might ask how we don't know that we should write $SL2$ as $SL1 + \Delta SL$? If we did this, we could eliminate $SL2$ and obtain the prediction equation

$$\hat{THEFT2} = 49.969 + .418 THEFT1 + .065 SL1 + .042 \Delta SL \tag{14}$$

In this equation, all coefficients are statistically significant, and, of course, R^2 remains unchanged at .231. This equation contains nothing out of the ordinary. It says that the initial level of social life and change in social life both slightly increase levels of involvement in theft. Alternately, we could eliminate $SL1$ and obtain the prediction equation

$$\hat{THEFT2} = 49.969 + .418 THEFT1 + .065 SL2 - .023 \Delta SL \tag{15}$$

Now the coefficient for the effect of change in social life on change in theft is of opposite sign to the contemporaneous effect of social life on theft. It is difficult to think of a theoretical reason why social life should increase theft, but an increase in social life should reduce theft. As it happens, the coefficient −.023 is not statistically significant ($p = .125$), and can be dropped from the equation. Had the sample size been larger, however, it would have been significant, and would then pose a problem in interpretation. In that circumstance, equation (15) would offer a less plausible understanding of the influence of social life on theft than equation (14).

It would be plausible, however, to think that change in social life leads to change in theft, with no instantaneous or lagged effect of a level of social life on time-2 theft. Estimating this model, with a lagged endogenous variable, we obtain

$$\hat{THEFT2} = 69.636 + .439 THEFT1 + .009 \Delta SL \tag{16}$$

In this equation, R^2 is .221, and the coefficient of .009 is not statistically significant ($p = .470$). The loss of explanatory power resulting from the omission of the contemporaneous effect in equation (15) is significant. We thus have clear evidence that the level of social life influences change in the level of theft, while a pure "change causes change" model doesn't perform.

Thus far we have been assuming that the model being estimated is one that directly describes the causal influences at work between the two variables. There are, however, additional circumstances in which it may be appropriate to estimate equation (10). Let us consider a few of the most commonly encountered examples.

Suppose that x influences y contemporaneously:

$$y_t = a + bx_t + e_t \tag{17}$$

with no lagged endogenous variable. Suppose also that the residuals, instead of being independent of one another, are characterized by first-order serial correlation,

$$e_t = \rho e_{t-1} + u_t \tag{18}$$

We assume that the u_t are independent and uncorrelated with $e_{t-1}$and with x_t. Although the OLS estimates of a and b in equation (17) are unbiased even in the presence of serially correlated errors, the estimates are inefficient.

[1] The data were collected through a complex sample design, making conventional inferential statistics invalid. Nevertheless, we follow the practice of earlier analysts of these data (Bynner, O'Malley and Bachman, 1981) in using standard t-tests, unadjusted for sample design. A proper analysis would use sample weights.

The standard errors are not as small as they would be in the absence of serially correlated errors. To obtain greater precision in estimation, we eliminate the serial correlation. To do this, multiply equation (17) by rho, and subtract the result from equation (17). After rearranging terms, we have

$$y_t = a(1-\rho) + \rho y_{t-1} - b\rho x_{t-1} + bx_t + u_t \qquad (19)$$

Now we have an equation that includes a lagged dependent variable, and has no serially correlated errors. If the model is correct, the coefficients are constrained; in the notation of equation (10),

$$-b_2/b_3 b_1 = 1 \qquad (20)$$

This constraint can be evaluated roughly by inserting estimates into equation (20) and seeing whether the equality is approximately valid. Taking the coefficients from Panel B of Table 17.3, we find that the left-hand side is −(.023)/(.042)/(.418) = −1.31. This is far from the right-hand side, suggesting that first-order serially correlated errors are not responsible for the presence of the lagged endogenous variable.

A more formal test of the constraint compares the goodness of fit when the constraint is imposed with the goodness of fit when all coefficients are estimated freely. If the restraints hold, the lagged endogenous variable may be due to serial correlation among the errors rather than being part of the causal model.

Next, consider a distributed lag model, in which x influences y with a strength that dies out gradually, instead of dissipating fully in just one time period:

$$y_t = a + b_1 x_t + b_2 x_{t-1} + b_3 x_{t-2} + \cdots$$
$$+ e_t + \Phi_1 e_{t-1} + \Phi_2 e_{t-2} \ldots \qquad (21)$$

Note that this equation, like equation (17), does not contain a lagged value of y. It predicts the level of y with no dynamic effects built in. However, it contains an infinite series in the lagged values of x. In addition, it is influenced by lagged shocks at past times, as well as at time t.

In empirical research, there can only be a finite number of waves, so equation (21) would have to be truncated in order to estimate it. In addition, the estimation of many coefficients eats up degrees of freedom. To avoid these problems, researchers often suppose that the influence of x on y diminishes geometrically with time and that the influence of the lagged shocks drops off in the same manner.[2] With this assumption, we can write equation (21) as

$$y_t = a + b_1(x_t + \lambda x_{t-1} + \lambda^2 x_{t-2} + \ldots)$$
$$+ e_t + \lambda e_{t-1} + \lambda^2 e_{t-2} \ldots \qquad (22)$$

If we lag this equation by one time unit, multiply it by lambda, subtract the result from equation (22), and rearrange terms, we obtain

$$y_t = a(1-\lambda) + \lambda y_{t-1} + b_1 x_t + e_t \qquad (23)$$

This procedure eliminates the infinite series; now we have only an instantaneous effect of x on y, but a lagged endogenous variable. Because there are now no serially correlated errors, this equation can be estimated by OLS regression without bias.

As these examples illustrate, a lagged endogenous variable can arise in a number of ways, and consequently the researcher should make an effort to differentiate these on the basis of both theoretical considerations and empirical evidence as to the constraints some of these possibilities imply. To complicate matters further, Allison (1990) has pointed out that in some circumstances the inclusion of a lagged dependent variable leads to counter-intuitive results, and that more meaningful results can be obtained by omitting it. These anomalies occur because

[2] Because omitted variables are the most common source of serially correlated errors, it is reasonable that this should be approximately true.

the presence of serially correlated errors in the true model leads to bias when not taken into account properly in the estimation.

4.3 Jointly dependent outcomes

Thus far we have been assuming that causal influences flow in just one direction – from *x* to *y* either contemporaneously, or with a lag, or both. Sometimes, however, there may be theoretical reasons for suspecting that *x* influences *y* and that *y* also influences *x*. There may also be reasons for thinking that any observed relationship between *x* and *y* is not causal but spurious. In circumstances like this panel data can help to clarify the relationships among variables.

For simplicity, consider the two-wave, two-variable model. The most general linear relationship among the interval-level variables *x* and *y* can be expressed by a pair of equations, one for each endogenous variable.

$$x_{i2} = a_1 + b_{11}x_{i1} + b_{12}y_{i1} + b_{13}y_{i2} + e_i$$
$$y_{i2} = a_2 + b_{21}x_{i1} + b_{22}y_{i1} + b_{23}x_{i2} + f_i \qquad (24)$$

In these equations *e* and *f* are random error terms. However, we do not assume that *e* is uncorrelated with *f*. This model is displayed visually in Figure 17.1, in which the Greek letter psi represents the covariance between the two error terms.

When causal influences flow in just one direction, from *x* to *y* but not from *y* to *x* (as in equation (10)), the solutions either explode or approach an equilibrium level asymptotically.

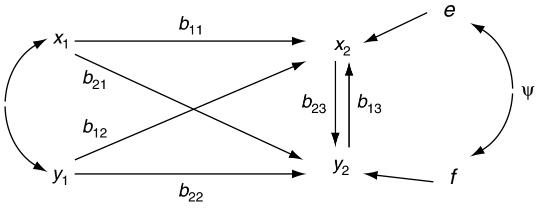


Figure 17.1 A two-wave, two-variable linear panel model

Where *x* and *y* influence one another reciprocally, the solutions can take on more complex forms. One can obtain solutions that explode, oscillate, or approach equilibrium monotonically or in an oscillatory manner. The expressions for the equilibrium values of *x* and *y* are complicated; they are given in Kessler and Greenberg (1979, pp. 118–22).

In many applications, the correlation between the two error terms, *e* and *f*, is most plausibly attributed to unobserved variables that influence both *x* and *y*. Consequently, the test for the significance of this correlation is also, in part, a partial test for spuriousness.

Estimation of equation (24) poses two problems. Because x_2 and y_2 mutually influence one another, ordinary least-squares estimates of the parameters in the equations suffer from simultaneity bias. To avoid this type of bias, the estimation should use instrumental variable methods (such as two-stage least squares) or maximum-likelihood estimation.

Second, in the absence of some additional information, equation (28) is under-identified. In the two-wave model for two standardized variables there are eight independent parameters to be estimated. If the variables are standardized, there are the two autoregressive coefficients, the two cross-lagged coefficients, the two cross-instantaneous coefficients, the correlation between the two time-one variables, and the correlation between the two prediction errors. Yet there are only six independent correlations among the four variables of the model available for carrying out the estimation. This means that it is impossible to obtain unique estimates of the parameters.

Most researchers achieve identification by assuming that the cross-instantaneous influences vanish, but it is also possible to assume that the cross-lagged influences vanish. If one makes either assumption, and allows the error terms *e* and *f* to be correlated, the model is just-identified. It will fit the variance–covariance matrix perfectly. Once one has a model in which the observed and theoretically predicted variances and covariances do not differ significantly, one can test individual coefficients for significance with *t*-tests or likelihood-ratio tests. To achieve a more parsimonious model, researchers often drop insignificant effects, always making sure that the overall goodness of fit does not deteriorate significantly.

To illustrate these methods, we continue to consider the relationship between juvenile delinquency and social life. We consider the possibility that each variable influences the other, either with a lag or contemporaneously. First, consider a model in which there are only lagged cross-effects, but no cross-contemporaneous causal effects. We do not constrain the correlation of prediction errors. The chi-square statistic for this model is zero, indicating a perfect fit to the data. Maximum-likelihood estimates of the standardized coefficients are shown in Panel A of Table 17.4, and below them, the *t*-statistic. The stability coefficients are both less than 1, so that regression to the mean is present for both variables. The cross-lagged influences are both positive and statistically significant, but the standardized coefficient for the effect of social life on theft is considerably greater than the coefficient for the effect of theft on social life. Psi, the estimate of the correlation between the two prediction errors, is close to zero and not statistically significant, indicating that omitted variables that influence both theft and social life are not significantly affecting the estimates.

We can modify the model so that the cross-influences between undesirable life events and theft are contemporaneous rather than lagged. With the correlation of prediction errors estimated from the data, this model also fits the data perfectly. The estimates for this model are in Panel B of Table 17.4. All stability coefficients are again less than 1, and the cross-influences are of comparable magnitude to those in Panel A. However, the correlation of prediction errors is significantly different from zero, suggesting that omitted variable bias or some other form of misspecification may be present.

Another way of testing for spuriousness is to consider models of the sort illustrated in Figure 17.2. In this model, the relationship between social life and involvement in theft is

Table 17.4 The influence on social life on juvenile theft in a two-wave panel model ($n = 1412$)

Independent variables	*Panel A*		*Panel B*	
	Dependent variables		*Dependent variables*	
	SL2	*THEFT2*	*SL2*	*THEFT2*
SL1	.499 (21.27)	.238 (10.24)	.463 (15.815)	
THEFT1	.050 (2.14)	.418 (17.96)		.367 (14.710)
SL2				.219 (2.120)
THEFT2			.262 (9.138)	
Ψ_{22}	−.018 (−.94)		−.459 (−7.03)	
R^2	.263	.275	.248	.089

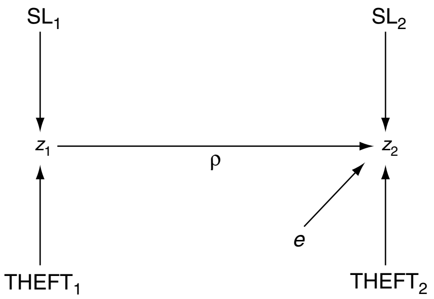


Figure 17.2 A two-wave, two-variable panel model of complete spuriousness

entirely spurious, due to the influence of an omitted variable z on both observed variables. This variable is assumed to be first-order autoregressive, i.e., $z_t = \rho z_{t-1} + e_t$. If one assumes that the coefficients a_j and b_j are constant over time, it follows from the rules of path analysis that the cross-lagged correlations of *SL1* with *THEFT2* and *THEFT1* with *SL2* will be equal. This is the Kenny criterion for spuriousness. However, without making that assumption, we see that another equality also holds:

$$r_{SL1,SL2}r_{THEFT1,THEFT2} = r_{SL1,THEFT2}r_{THEFT1,SL2} \quad (26)$$

Inserting the observed values for these correlations, we see that neither inequality is close to being obeyed. The estimate of left-hand side in equation (25) is .33; the right-hand side is .16. The left-hand side of equation (26) is (.51)(.47) = .24; the right-hand side is (.33)(.16) = .05. The chi-square test for the model as a whole is 218.41, with one degree of freedom. This is highly significant, indicating that the fit is grossly unacceptable. Consequently, we conclude that the relationship between social life and theft is not entirely spurious. However, we cannot exclude the possibility that it is *partly* spurious. Consequently, the estimates for the causal models must be interpreted with caution; they are predicated on the model itself being correctly specified.

We can now return to the comparison of cross-lagged correlations to see the superiority of structural equation modeling. To test for the significance of the difference between the correlation of *SL1* with *THEFT2* and *THEFT1* with *SL2*, one would compute the Pearson-Filon statistic (Kenney, 1979, pp. 238–39), and discover that the statistic in this case is statistically significant. One would then conclude that social life influences theft more than theft influences social life. A structural equation modeler can carry out the same test by constraining the two cross-influences to be the same and comparing the chi-square statistics of the constrained and unconstrained models. Because the two models are nested (one is a special case of the other), the difference in chi-square statistics is itself a chi-square statistic, one which, in the present case, is highly significant (chi-square = 14.82 with one degree of freedom, p = .00012), We thus reach the same conclusion as would be obtained from the CLPC approach, but we learn a great deal more. In particular, we learn that both cross-influences are statistically significant, something that would not be learned in a CLPC analysis. In addition, we obtain numerical estimates of the parameters in the model.

4.4 Pooling methods

Pooling methods offer a different way of dealing with the question of spuriousness (Johnston and DiNardo, 1997, pp. 388–409; Greene, 2000; Wooldridge, 2001; Baltagi, 2005; see also Worrall, Chapter 15 and Finkel, Chapter 29 in this volume). Consider this equation for the influence of x at time t on y at the same time. In recognition of the possibility that additional unmeasured variables may influence y we add to the conventional regression equation a term α_i that is constant over time. We can then write our regression equation as

$$y_{it} = a + bx_{it} + \alpha_i + e_{it} \quad (27)$$

The term in alpha introduces unmeasured heterogeneity into the equation. It is tantamount to introducing a dummy variable for every case into the regression equation. This term gives each case its own intercept. The slope, however, is taken to be the same for all cases. We make the conventional assumptions about the e_{it}: they are normally distributed with a mean of zero, and are uncorrelated with x_{it}. They are independent and identically distributed.

If we only had cross-sectional data, we could not estimate equation (27) without additional assumptions. With panel data, however, several estimation strategies are possible. It is tempting simply to “stack” the data, so that all the observations at time $t = 1$ come first, followed by all the observations at time 2, then all those at time 3, and so forth. If there are n cases and T waves, this would mean NT observations.

However, because the same cases are observed at each wave, the observations are not independent. Statistically speaking, the error term for a given case at a given time is

$$u_{it} = \alpha_i + e_{it} \tag{28}$$

The common presence of alpha will create a positive correlation between the values of u at different times. This violates the OLS assumption that residuals are independent.

One way to estimate the parameters that does not ignore this complication is to assume that the alphas are uncorrelated with the x_{it}. When this assumption is made, we have a *random effects model*. The coefficient b can be estimated by means of feasible generalized least squares, a procedure that utilizes information about variation between cases and over time. The estimates are consistent, that is to say, unbiased as the sample size grows without limit. In this procedure we do not estimate the individual effects (the α_{i}).

An important limitation to this method is that it will yield biased estimates of b if the α_{i} are in fact correlated with the x_{it}. Most of the time we don’t know what variables have been omitted, and consequently have no theoretical grounds for assuming that they are uncorrelated.

An alternative approach is to estimate a *fixed effects model*. We evaluate equation (27) at the mean of each variable (the mean is taken over time)

$$y_{it} = \overline{a} + \overline{bx}_{it} + \overline{\alpha}_i + \overline{e}_{it} \tag{29}$$

The means of a and α_{i} are, of course, just a and α_i, respectively. Subtracting equation (29) from equation (27), we have

$$(y_i - \overline{y}_i) = b(x_{it} - \overline{x}_i) + (e_{it} - \overline{e}_i) \tag{30}$$

The constant term and the individual effects have all dropped out. Moreover, the observations are independent of one another across time. This equation can be estimated by OLS regression. This estimate uses only information about the variation in scores over time; it doesn’t make use of whatever information is contained in the cross-sectional variation. The estimates of b will be consistent provided that *strict exogeneity* holds. This means that x_{it} is uncorrelated with the residuals not just at time t, but at all times, past and present. This requirement is quite stringent; it precludes long-term influences of y on x. In many applications, this assumption is dubious. The precision with which the estimates of the α_{i} are estimated does not increase with the sample size, but it does increase with the number of waves.

An important disadvantage of the fixed effect model is that any time-invariant observed variables will drop out in the subtraction. Consequently, this method will provide no information about the effects of stable personality traits, regional subcultures and the like, even if these variables have been measured. This can be a major liability. Mover, the fixed effects estimator is less efficient than the random effects estimator. Consequently, one might prefer to use the random effects estimator, if one could feel confident that the α_{i} were indeed uncorrelated with the independent variables.[3] The Wu-Hausman test (see, e.g., Hausman, 1978) can be used to determine whether there are significant differences between the estimates obtained from the two procedures.

It is common when carrying out either method of estimation, to introduce fixed effects for each wave (except the first). This controls for trends in the outcome that are common to all cases. For example, in a study of American states, this would control for developments that affect all states equally. These fixed effects can be introduced by adding dummy variables for each wave but one to the model.

These methods are illustrated by Phillips and Greenberg's (2005) study of the factors that influence homicide rates in 400 American counties during the years 1985–1999. Fixed effects and random effects estimators for the effects of 11 independent variables obtained using Stata version 9.1 are shown in Table 17.5. Note that some of the coefficients are significant in the random effects model but not in the fixed effects model. The population variables are significant in both models, but with opposite signs. These differences make it important to know which model to believe. A Wu-Hausman test prefers the fixed effects model. Evidently some of the effects that are significant in the random effects estimation are actually spurious.

As conventionally implemented, these methods do not furnish the overall goodness-of-fit test that structural equation modeling provides. Recently, Paul Allison (2005) has shown how the fixed effects model can be cast in a SEM framework, so that the diagnostic indicators that SEM provides can be used here as well.

A complication, until recently given little attention in discussion of panel models, concerns stationarity. It has long been known in time series methodology that conventional statistical methods break down if the data being analyzed are not stationary. Stationarity holds when means, variances, and covariances are constant over time. Trends constitute one commonly encountered form of nonstationarity. When present they can lead to spurious estimates. Nonstationarity can also occur in panel data. There are now tests for it, and special methods for handling it when it occurs.

The basic approach just outlined can easily be modified to accommodate lagged influences and serially-correlated errors. If there are reciprocal influences, so that some of the independent variables are endogenous, or if one of the predictors is a lagged endogenous variable (a dynamic panel model), special estimation techniques (instrumental variables) are needed.

A different estimation technique is also needed when the number of cases in the data set is fairly small. Though the random effects estimator may be consistent, its small-sample properties are not well understood. Simulations carried out by Nathaniel Beck and Jonathan Katz (1995) demonstrate that the standard deviations obtained in this way are much too small. They propose to use OLS estimates with standard errors corrected to reflect the panel structure of the data. Their *panel-corrected standard errors* approach performs much better in simulations.

By writing the equations being pooled with an error term that is taken to be normally distributed, we are implicitly saying that the dependent variable is also continuous. Pooling models can also be used when the dependent variable is categorical, ordinal, or count. However, there is no fixed effect estimator for analyzing logits and probits from panel data, because there is no way to eliminate the fixed effects.

4.5 Latent growth curve modeling

Researchers studying human development have developed still another method for carrying out causal analysis with panel data – latent

[3] If that assumption is untenable, an instrumental variables estimation is still possible.

Table 17.5 Fixed effects and random effects models of logged homicide rate

Variable	*Fixed effects model*		*Random effects model*	
	Coefficient	*Standard Error*	*Coefficient*	*Standard Error*
Intercept	0.811	1.225		
% divorced	0.023	0.013#	0.104	0.008*
Unemployment	−0.005	0.004	0.007	0.004
Per capita income (000$)	−0.004	0.007	−0.011	0.004*
% aged 15–34	0.036	0.008*	0.010	0.004*
% male	−0.008	0.026	−0.046	0.014*
% black	0.024	0.007*	0.038	0.002*
Population (000,000)	−0.667	0.120*	0.245	0.046*
Population^2	0.051	0.018*	−0.016	0.007*
West			0.298	0.054*
South			0.255	0.045*
Northeast			0.065	0.046
R-square				
Within counties	0.086		0.070	
Between counties	0.203		0.809	
Overall	0.184		0.692	
Fraction of variance due to fixed effects	0.814		0.457	

Note: Both models include fixed effects for years, but the estimates are omitted from the table.
# Significant at the .10 level
* Significant at the .05 level

growth curve models. This approach postulates a "level-one" model that characterizes the overall growth pattern in the data set. It takes the form of a linear or quadratic equation in time, with a random error term:

$$y_{it} = a_i = b_{1i}t + b_{2i}t^2 + e_{it} \tag{31}$$

We make the standard assumptions about the residuals. What is distinctive about this equation is that the coefficients have subscripts. Latent growth curve models allow the intercepts and slopes to vary from case to case. This extends the assumption in pooling methods that intercepts, but not slopes, can vary from case to case.

Variables that influence the intercepts and slopes are specified in a "level-two" model, such as

$$\begin{aligned} a_i &= \gamma_{00} + \gamma_{01}x_{it} + u_{1i} \\ b_{1i} &= \gamma_{10} + \gamma_{11}x_{it} + u_{2i} \\ b_{2i} &= \gamma_{20} + \gamma_{21}x_{it} + u_{3i} \end{aligned} \tag{32}$$

The coefficients in the level-two equations for the slopes (b_1 and b_2) can be interpreted as interaction effects. They indicate that the time dependence of the growth curve varies systematically (nonrandomly) across cases, or, equivalently, that the effect of x on y depends on time. Additional explanatory variables can be added

Table 17.6 Latent growth curve model of logged homicide rates

Variable	*Intercept*		*Linear term*		*Quadratic term*	
	Coeff.	*SE*	*Coeff.*	*SE*	*Coeff*	*SE*
Intercept	1.398	0.852	0.223	0.206	0.001	0.014
% divorced	0.108	0.013*	0.001	0.003	0.0001	0.000
Unemployment	0.018	0.007*	0.001	0.002	−0.0002	0.000
Per capita income (000$)	−0.005	0.008	−0.003	0.002#	0.0002	0.000
% aged 15–34	0.006	0.006	−0.001	0.001	0.0001	0.000
% male	−0.028	0.018	−0.003	0.004	−0.0001	0.000
% black	0.028	0.002*	0.003	0.000*	−0.0002	0.000*
Population (000,000)	0.450	0.065*	−0.006	0.014	−0.0008	0.001
Population^{2}	−0.043	0.010*	0.002	0.002	0.0000	0.000
West	0.322	0.064*	−0.024	0.015	0.0018	0.001#
South	0.521	0.054*	−0.050	0.012*	0.0013	0.001
Northeast	0.051	0.058	−0.029	0.014*	0.0013	0.001
−2 log likelihood	4233.2					

Note: # = significant at the .10 level;
∗ = significant at the .05 level.

at will. Moreover, one can construct a "level-three" equation in which the coefficients of the "level-two" equation depend on a set of still other variables. For example, in studying test scores of schoolchildren, the level-two model might specify attributes of the child that influence the level and rate of improvement in scores over time, while the level-three equation might allow these influences to depend on characteristics of the school or neighborhood.

A comparison with the pooling approach is instructive. The latent growth curve model specifies a particular functional dependence on time, while the pooled model, by using dummies for each time, does not impose any temporal dependence on the model.[4] As usually employed, the latent growth curve model does not incorporate fixed effects for unmeasured stable attributes of cases, though this can be done. The pooled model does not have interaction terms between the explanatory variables and time – though these can be added to the model should one wish to do so.

Analyzing the county homicide data in this way, Phillips and Greenberg (2005) obtained the maximum-likelihood estimates shown in Table 17.6 using SAS version 9.1. Comparison of the two sets of estimates shows some similarities – e.g, *percent divorced* and *percent black* both have significantly positive effects on homicide rates in both models. Other variables, however, are significant in one model but not the other. For example, *unemployment* fails to influence homicide rates significantly in the pooled model, but has a significantly positive effect in the growth curve model. These differences reflect differences in the assumptions each model makes about the processes generating homicide rates.

As implemented in software packages, routines for doing this type of estimation do not readily accommodate models in which there are mutually dependent variables. However, this type of estimation can be done easily when the

[4] The growth curve approach can accommodate other types of time dependence, but this is rarely done.

multilevel model is recast as a structural equation model (Bollen and Curran, 2005).

5 Independent cross-sections

Sometimes panel data are not available, but repeated cross-sectional data (independent cross-sections at more than one time) are. Because the same cases do not appear at each time, it is impossible to take differences. Though aggregate change can be observed, one doesn't know which cases in a sample have changed. This poses limits to the types of causal analysis that can be conducted. Obviously, fixed effects models cannot be estimated. Nevertheless, it is possible to carry out causal analyses. The independence of observations means that we can legitimately stack the data and ignore the fact that we have multiple waves when we estimate our regression equations.

A particular technique, "difference-in-difference", is particularly useful for analyzing the effects of discrete events from independent cross-sections. Suppose that we have observations for dependent variable y both before and after an event of some sort happens to some, but not all of the cases. The event could be a planned intervention, such as the adoption of a new law or policy; or it may be an unplanned, exogenous event such as a natural disaster. The variable Z represents the innovation. We code it 0 for cases to which the event did not happen, and 1 for those that experienced the event.

A naive approach would be to estimate a regression equation of the form

$$y_i = a + bZ_i + e_i \tag{33}$$

using just the post-event data. Doing this we could estimate $\hat{a} = \overline{y}$ using only data for cases that did not experience the event, and $\hat{a} + \hat{b} = \overline{y}$ for the cases that did. By subtraction we obtain an estimate of b, which expresses the impact of the event. This procedure, however, fails to take into account differences between the cases that may have existed before the innovation was introduced. Consequently, the possibility of spuriousness cannot be ruled out.

An alternative naive strategy would be to compare levels of y in adopting locations before and after the innovation was adopted. The estimating procedure would be essentially the same, but this time the same cases would be compared at two times. This approach would control for pre-existing differences among the locations, but would not exclude the possibility that the changes are due to some other cause than the new policy.

The "difference-in-difference" strategy controls for both possibilities (Ashenfelter and Card, 1985; Abadie, 2005). Calling the time before any events occurred $t = 0$, and those after they occurred $t = 1$, we write the regression equation as

$$y_{it} = a_0 + a_{1i} + a_2 t + bZ_{it} + e_{it} \tag{36}$$

The first coefficient represents a contribution to the outcome that is the same for all places at both times. The second represents factors that differ from one case to another but are constant over time. They might, on average, be different in places that fail to adopt a new policy, than in places that do, or in families that undergo a divorce, and in those that do not. The third coefficient represents trends that are common to all places. The coefficient b continues to represent the effect of the event.

Taking the mean of this equation at each time and evaluating it when $t = 0$ and 1, and when $Z = 0$ and 1, we obtain four equations that can be solved for the four parameter estimates. The estimate for b is

$$\begin{aligned}\hat{b} = {} & [\overline{y}(t=1, Z=1) - \overline{y}(t=0, Z=1)] \\ & - [\overline{y}(t=1, Z=0) - \overline{y}(t=0, Z=0)]\end{aligned} \tag{37}$$

This is the "difference in difference". It is the difference between the change in the dependent variable for cases that experienced the event, and the change for cases that did not. It can also be expressed as the difference in outcomes between the two sets of cases after the events occurred, less the difference in outcomes before any events occurred.

The method can be broadened to include additional explanatory variables, and to circumstances in which the innovation variable is not dichotomous, but the level of a quantitative variable. In the form presented here, the method implicitly assumes that in the absence of any exogenous events, the two sets of cases would undergo the same temporal change in the dependent variable. With more than two waves, each set of cases can be allowed to have its own inherent growth rate (Abadie, 2005).

In some contexts, the implications of an event should be different for different groups of people. For example, civil rights laws barring racial discrimination in employment should have different impacts on employment for white and black workers. To study the impact of state laws barring racial discrimination prior to the adoption of a federal civil rights law, Collins (2003) looked for such differences using a "difference-in-difference-in difference" analysis. His analysis first looked at before-and-after changes in employment and wages for states that adopted laws barring racial discrimination as compared with states that did not. These differences were computed separately for white and black workers, and the differences differenced.

6 Software

The types of analysis discussed under Section 3 Qualitative Outcomes can be done with virtually any statistical software package. This is not true for the analyses described under Quantitative Outcomes. Pooled methods are most easily implemented with packages designed for this purpose, e.g., Stata, SAS, Limdep and Eviews. Structural equation modeling can be carried out with such specialized programs as LISREL, Amos, EQS and Mplus. It can also be done in SAS, using PROC CALIS. Multi-level modeling can be carried out in special stand-alone programs such as HLM and MlwiN, and in some general statistical packages, including SPSS, Stata and SAS.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72: 1–19.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. In C. Clogg (ed.), *Sociological Methodology*, pp. 93–114, Volume 20. Boston: Blackwell.

Allison, P. D. (2005). *Fixed Effects Regression Models for Longitudinal Data Using SAS*. Cary, NC: SAS Press.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effects of training programs. *Review of Economics and Statistics*, SE 67: 684–60.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*, 2nd edn. Wiley: New York.

Beck, N. and Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, 89: 634–47.

Bollen, K. and Curran, P. (2005). *Latent Curve Models: A Structural Equation Perspective*. New York: Wiley.

Bynner, J. M., O'Malley, P. M. and Bachman, J. G. (1981). Self-esteem and delinquency revisited. *Journal of Youth and Adolescence*, 10: 407–41.

Campbell, D. and J. S. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally.

Coleman, J. S. (1964). *Introduction to Mathematical Sociology*. New York: Free Press.

Collins, W. J. (2003). The labor market impact of state-level antidiscrimination laws, 1940–1960. *Industrial Labor Relation*, 56: 244–72.

Goodman, L. (1973a). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach, *Biomet*. 60: 179–92.

Goodman, L. (1973b). Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology*, 78: 1135–91.

Greene, W. H. (2000). *Econometric Analysis*, 4th edn. Upper Saddle River, NJ: Prentice-Hall.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46: 1251–1271.

Johnston, J. and DiNardo, J. (1997). *Econometric Methods*, 4th edn. New York: McGraw-Hill.

Kenny, D. A. (1975). Cross-lagged panel correlation: A test for spuriousness, *Psychology Bulletin*, 82: 345–62.

Kenny, D. A. (1979). *Correlation and Causality*. New York: Wiley.

Kessler, R. C. (1977). Rethinking the 16-fold table, *Social Science Research*, 6: 84–107.

Kessler, R. C. and Greenberg, D. F. (1979). *Linear Panel Analysis: Models of Quantitative Change*. New York: Academic Press.

Lazarsfeld, P. F. (1972) Mutual effects of statistical variables. In P. F. Lazarsfeld, A. K. Pasanaella and M. Rosenberg (eds), *Continuities in the Language of Social Research*. New York: Free Press.

Lazarsfeld, P. F. and Fiske, M. (1938). The "panel" as a new tool for measuring opinion. *Public Opinion Quarterly*, 2: 596–612.

McCullough, B. C. (1978). Effects of variables using panel data: A review of techniques. *Public Opinion Quarterly*, 42: 199–220.

Miller, T. Q. and Flay, B. R. (1996). Using loglinear models for longitudinal data to test alternative explanations for stage-like phenomena: An example from research on adolescent substance use. *Mult. Behav. Res.*, 31: 169–96.

Pelz, A. and Andrews, F. M. (1964). Detecting causal priorities in panel study data. *American Socological Review*, 29: 836–48.

Phillips, J. A. and Greenberg, D. F. (2005). A comparison of methods for analyzing criminological panel data. Paper presented to the American Society of Criminology.

Rosenberg, F. R. and Rosenberg, M. (1978). Self-esteem and delinquency. *Journal of Youth and Adolescence*, 7: 279–94.

Wooldridge, J. M. (2001). *The Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT Press.

Chapter 18

Causal inference in longitudinal experimental research

Jos W. R. Twisk

1 Introduction

Randomized controlled trials (RCTs) are considered to be the gold standard for evaluating the effect of a certain intervention (Rothman and Greenland, 1998). In a randomized controlled trial, the population under study is randomly divided into an intervention group and a nonintervention, or reference, group (e.g., a placebo group or a group with "usual" care). When discussing the analysis of causal inference in experimental research a distinction must be made between studies with only one follow-up measurement and studies with more than one follow-up measurement. When there is only one follow-up measurement relatively simple statistical techniques can be used to evaluate the effect of the intervention, while when more than one follow-up measurement is considered, in general, more sophisticated statistical techniques are necessary.

2 Experimental research with only one follow-up measurement

In most RCTs, besides the follow-up measurement a baseline measurement is also performed. With the information gathered in this baseline measurement, it is possible to compare the changes in the outcome variable between the intervention and the reference group. Although this procedure looks quite straightforward, the definition of change can be complicated. In fact, since the beginning of the 1960s, there has been an ongoing debate how to define "change" (Bereiter, 1963; Cronbach and Furby, 1970; Plewis, 1985; Gottman, 1995). When evaluating the RCT literature, in most research situations the absolute change between a baseline measurement and a follow-up measurement is calculated, and this absolute change in a certain outcome variable is compared between the groups of interest. However, there are many ways to define changes between a baseline and a follow-up measurement (Twisk, 2003), and therefore there are many ways to evaluate the results of a (randomized controlled) trial.

2.1 Changes between baseline and follow-up: continuous outcome variables

As mentioned before, the simplest method is to calculate the absolute difference between two measurements over time [equation (1)].

$$\Delta Y = Y_{it2} - Y_{it1} \qquad (1)$$

where: Y_{it2} = observations for subject i at time $t2$; and Y_{it1} = observations for subject i at time $t1$.

One of the typical problems related to the use of absolute change in RCTs to evaluate the effect of an intervention, is the phenomenon of regression to the mean. If the outcome variable at $t = 1$ is a sample of random numbers, and the outcome variable at $t = 2$ is also a sample of random numbers, then the subjects in the upper part of the distribution at $t = 1$ are less likely to be in the upper part of the distribution at $t = 2$, compared to the other subjects. In the same way, the subjects in the lower part of the distribution at $t = 1$ are less likely than the other subjects to be in the lower part of the distribution at $t = 2$. The consequence of this is that, just by chance, the change between $t = 1$ and $t = 2$ is correlated with the initial value. Another consequence is that when the intervention group and control group differ at baseline, a comparison of the absolute changes between the groups can lead to either an overestimation or an underestimation of the intervention effect (Twisk and Proper, 2004).

There are, however, some ways in which it is possibile to define changes between subsequent measurements, more or less "correcting" for the phenomenon of regression to the mean. One of these possibilities is the use of a relative difference between two measurements over time instead of the absolute difference [equation (2)]

$$\Delta Y = \frac{(Y_{it2} - Y_{it1})}{Y_{it1}} \times 100\% \qquad (2)$$

where: Y_{it2} = observations for subject i at time $t2$; and Y_{it1} = observations for subject i at time $t1$.

Although it has been suggested that the use of relative change corrects for the phenomenon of regression to the mean, this is not the case. Figure 18.1 illustrates this artefact.

Figure 18.1 shows the development over time in a certain continuous outcome variable for an intervention and a control group. In the first part of the figure, the intervention group decreases from 4 to 3 (i.e., a 25% decrease), while the control group decreases from 3 to 2 (i.e., a 33% decrease). Using the relative change in this situation will more or less correct for the phenomenon of regression to the mean, because the change in the control group is "more difficult" and therefore an equal absolute change will be evaluated in favor of the control group. However, when the outcome variable increases over time, which is illustrated in the second part of Figure 18.1. The intervention group increases from 3 to 4 (i.e., a 33% increase), while the control group increases from 2 to 3 (i.e., a 50% increase). Using the relative change in this situation will not correct for the phenomenon of regression to the mean. In fact, it will exacerbate the problem. Another approach in which to define changes correcting for regression to the mean is known as "analysis of covariance" [equation (3)]. With this technique the value of the outcome variable Y at the second measurement is used as the outcome variable in a linear regression analysis, with the observation of the outcome variable Y at the first measurement as one of the predictor variables (i.e., as a covariate):

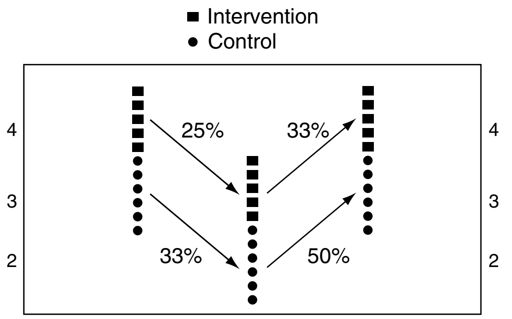


Figure 18.1 Relative change does not always correct for the phenomenon of regression to the mean

$$Y_{it2} = \beta_0 + \beta_1 Y_{it1} + \beta_2 X_i + \ldots\ldots. \qquad (3)$$

where: Y_{it2} = observations for subject i at time $t2$; β_1 = regression coefficient for Y_{it1}; Y_{it1} = observations for subject i at time $t1$; β_2 = regression coefficient for X_i; and X_i = intervention variable.

In the analysis of covariance, the change is defined *relative* to the value of Y at $t=1$. This relativity is expressed in the regression coefficient β_1 and, therefore, it is assumed that this method 'corrects' for the phenomenon of regression to the mean. In fact the effect of the intervention is evaluated assuming the same baseline value for both groups. Some researchers argue that the best way to define changes, correcting for the phenomenon of regression to the mean, is a combination of equations (1) and (3). They suggest calculating the absolute change between Y_{t2} and Y_{t1}, correcting for the value of Y_{t1} (equation 4).

$$Y_{it2} - Y_{it1} = \beta_0 + \beta_1 Y_{it1} + \beta_2 X_i + \ldots\ldots \tag{4}$$

where: Y_{it2} = observations for subject i at time $t2$; Y_{it1} = observations for subject i at time $t1$; β_1 = regression coefficient for Y_{it1}; β_2 = regression coefficient for X_i; and X_i = intervention variable.

However, analyzing the change, correcting for the initial value at $t=1$, is exactly the same as the analysis of covariance described in equation (3). This can be seen when equation (4) is written in another way [equation (5)]. The only difference between the models is that the regression coefficient for the initial value is different; i.e., the difference between the regression coefficients for the initial value is equal to 1.

$$Y_{it2} = \beta_0 + (\beta_1 + 1)\ Y_{it1} + \beta_2 X_i + \ldots\ldots \tag{5}$$

where: Y_{it2} = observations for subject i at time $t2$; β_1 = regression coefficient for Y_{it1}; Y_{it1} = observations for subject i at time $t1$; β_2 = regression coefficient for X_i; and X_i = intervention variable.

All techniques discussed so far are suitable in situations in which the continuous outcome variable theoretically ranges from 0 to +∞, or from −∞ to 0, or from −∞ to +∞. Some variables (e.g., scores on questionnaires) have maximal possible values ("ceilings") and/or minimal possible values ("floors"). To take these "ceilings" and/or "floors" into account, the definition of change can be as shown in equation (6).

$$\text{when } Y_{it2} > Y_{it1} : \Delta Y = \frac{(Y_{it2} - Y_{it1})}{(Y_{\max} - Y_{it1})} \times 100\% \tag{6a}$$

$$\text{when } Y_{it2} < Y_{it1} : \Delta Y = \frac{(Y_{it2} - Y_{it1})}{(Y_{it1} - Y_{\min})} \times 100\% \tag{6b}$$

$$\text{when } Y_{it2} = Y_{it1} : \Delta Y = 0 \tag{6c}$$

where: Y_{it2} = observations for subject i at time $t2$; Y_{it1} = observations for subject i at time $t1$; Y_{max} = maximal possible value of Y ("ceiling"); and Y_{min} = minimal possible value of Y ("floor").

2.2 Changes between baseline and follow-up: dichotomous (and categorical) outcome variables

For dichotomous outcome variables the situation is slightly more difficult than was described for continuous outcome variables. This is due to the fact that a change in a dichotomous outcome variable between subsequent measurements leads to a categorical variable. First of all, there are subjects who stay in the "highest" category, there are subjects who stay in the "lowest" category, and there are subjects who move from one category to another (see Figure 18.2).

In general, for categorical outcome variables with C categories, the change between subsequent measurements is another categorical variable with C^2 categories. The cross-sectional analysis of the resulting categorical variable can be performed with polytomous/multinomial logistic regression analysis, which is now available in most software packages.

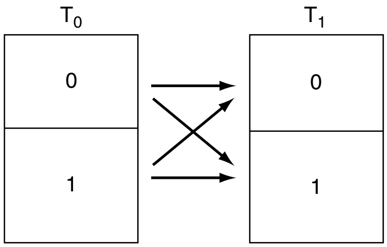


Figure 18.2 Changes in a dichotomous outcome variable over time results in a categorical outcome variable

Unfortunately, polytomous logistic regression analysis is not much used. Therefore, in many studies the resulting categorical outcome variable is reduced to a dichotomous outcome variable, which can be analyzed with simple logistic regression analysis. One widely used possibility is to discriminate between subjects who showed an "increase" and subjects who did not, etc. Nevertheless, in every reduction information is lost, and it is obvious that a dichotomization is not recommended in most research situations.

Another method with which to analyze changes in a dichotomous outcome variable is analysis of covariance. Instead of a linear regression analysis, logistic regression analysis must be used for the dichotomous outcome variable.

For dichotomous outcome variables, the interpretation of the results of a logistic analysis of covariance is, however, complicated, especially when all four possible changes over time occur in the dataset. (For further detail on the use of logistic regression to analyze changes in dichotomous and categorical outcomes, see the chapters by Menard in Section VI of this volume.)

2.3 Changes between baseline and follow-up: ordinal outcome variables

The definition of change between baseline and follow-up is even more complicated when ordinal outcome variables are considered. Ordinal scales are characterized not only by an ordering of categories but also by the fact that the distance from category to category is not known. In other words, on ordinal scales, the difference between 2 and 3 cannot be assumed the same as the difference between 3 and 4 (Stucki et al., 1996). This implies that analyses of changes in ordinal outcome variables are only valid when there are no differences at baseline. When there are differences at baseline, none of the earlier presented methods leads to valid results. When an ordinal outcome variable is based on a sum score of several items of a questionnaire (which is common in health-status measures), Rasch analysis (Raczek et al., 1998; MacKnight and Rockwood, 2000) can be seen as a sort of recalibration of the ordinal scale into an interval scale. In general, it is suggested to use nonparametric statistics to analyze changes in ordinal outcome variables (Schnell et al., 1995; Sonn and Svensson, 1997; Svensson, 1998), or to use complicated statistical modeling (Agresti, 1989; 1999). However, whatever the statistical technique used to analyze changes in ordinal outcome variables, the results of these analyses should be interpreted cautiously.

2.4 Recommendation

Although in most studies the absolute change between a baseline measurement and a follow-up measurement is used to evaluate the effect of certain intervention, in many situations this is not the most appropriate method, first of all because of its assumed negative correlation with the initial value (i.e., the phenomenon of regression to the mean), and second because of its low reliability. For more information on the latter, reference is made to Rogosaa (1995) who gives an interesting overview of the "myths and methods" in longitudinal research and, in particular, the definition of change.

It is difficult to give straightforward advice regarding the definition of change that should be used in an RCT with two measurements. The choice for a particular method depends on the characteristics of the outcome variable. When a continuous outcome variable is involved and there are no anticipated ceiling or floor effects, analysis of covariance is recommended because the technique corrects (if necessary) for the phenomenon of regression to the mean. When there are anticipated ceiling or floor effects, they should be taken into account. However, the best way of analyzing change is (probably) a combination of the results of several analyses obtained from various (biologically plausible) definitions of change. When changes in a dichotomous outcome variable are analyzed, polytomous logistic regression analysis of the categorical variable is preferable to (logistic) analysis of covariance, because it provides more information and the interpretation of the results is fairly straightforward.

3 Experimental research with more than one follow-up measurement

In the past decade, experimental studies with only one follow-up measurement have become very rare. At least one short-term follow-up measurement and one long-term follow-up measurement "must" be performed. However, more than two follow-up measurements are usually performed in order to investigate the "development" of the outcome variable, and to compare the "developments" among the groups. These more complicated experimental designs are often analyzed with the simple methods that have already been described in the earlier paragraphs, mostly by analyzing the outcome at each follow-up measurement separately, or sometimes even by ignoring the information gathered from the in-between measurements, i.e., only using the last measurement as an outcome variable to evaluate the effect of the intervention. Besides this, summary statistics are often used (see Section 3.1). This is surprising, because there are statistical methods available that can be used to analyze the difference in "development" of the outcome variable in two or more groups.

3.1 Continuous outcome variables

Summary statistics

There are many summary statistics available with which to estimate the effect of an intervention in an experimental study. In fact, the relatively simple analyses carried out in Section 2 can also be considered as summary statistics. Depending on the research question to be addressed and the characteristics of the outcome variable, different summary statistics can be used. The general idea of a summary statistic is to express the longitudinal development of a particular outcome variable as one quantity. Therefore, the complicated longitudinal problem is reduced to a cross-sectional problem. To evaluate the effect of the intervention, the summary statistics of the groups under study are compared to each other. Table 18.1 gives a few examples of summary statistics.

One of the most frequently used summary statistics is the area under the curve [AUC; equation (7)].

$$\text{AUC} = \frac{1}{2}\sum_{t=1}^{T-1}(t_{t+1} - t_t)(Y_t + Y_{t+1}) \tag{7}$$

where: AUC = area under the curve; T = number of measurements; and Y = observation of the outcome variable at time = t.

When the AUC is used as a summary statistic, the AUC must first be calculated for each subject, which is then used as an outcome variable to evaluate the effect of the therapy under study. This comparison is simple to carry out with an

Table 18.1 Examples of summary statistics, which are frequently used in experimental studies

- The mean of all follow-up measurements
- The highest (or lowest) value during follow-up
- The time needed to reach the highest value or a certain pre-defined level
- Changes between baseline and follow-up levels
- The area under the curve

independent *t*-test. When the time-intervals are equally spaced, the AUC is exactly the same as the overall mean. The AUC becomes interesting when the time-intervals in the longitudinal study are unequally spaced, because then the AUC reflects the "weighted" average in a certain outcome variable over the total follow-up period.

(M)ANOVA for repeated measurements

The basic idea behind (multivariate) analysis of variance ((M)ANOVA) for repeated measurements (which is also known as a "general linear model for repeated measurements") is the same as for the well-known paired *t*-test. The statistical test is carried out for the *T*-1 absolute differences between subsequent measurements. In fact, (M)ANOVA for repeated measurements is a multivariate analysis of these *T*-1 absolute differences between subsequent time-points. Multivariate refers to fact that *T*-1 differences are used simultaneously as outcome variables. Besides the "multivariate" approach, the same research question can also be answered with a "univariate" approach. This "univariate" procedure is comparable to the procedures carried out in simple analysis of variance (ANOVA) and is based on the "sum of squares", i.e., squared differences between observed values and average values. In most software packages, the results of both the "multivariate" and "univariate" approaches are provided at the same time. From a (M)ANOVA for repeated measurements with one dichotomous determinant (i.e., intervention versus control group), basically three "effects" can be derived (Twisk, 2003). An overall time-effect (i.e., is there a change over time, independent of the different groups), an overall group effect (i.e., is there a difference between the groups independent of time) and, most important, a group-time interaction effect (i.e., is there a difference between the groups in development over time). Although (M)ANOVA for repeated measurements is very often used, it has a few major drawbacks. First of all, it can only be applied to complete cases; all subjects with one or more missing observation are not part of the analyses. Second, (M)ANOVA for repeated measurements is mainly based on statistical significance testing, while there is more interest in effect estimation. Because of this, nowadays sophisticated statistical techniques, such as generalized estimating equations (GEE) or random coefficient analysis (see below), are becoming more and more popular.

(M)ANOVA for repeated measurements corrected for the baseline value

When the baseline values are different in the groups to be compared, it is often suggested that a (M)ANOVA for repeated measurements should be performed, correcting for the baseline value of the outcome variable. It should be noted carefully that when this procedure (which is also known as (multivariate) analysis of covariance, i.e., (M)ANCOVA) is performed the baseline value is both an outcome variable (i.e., to create the difference between the baseline value and the first follow-up measurement) and a covariate. In some software packages (such as SPSS) this is not possible, and therefore an exact copy of the baseline value must be added to the model.

Sophisticated analysis

The questions answered by (M)ANOVA for repeated measurements could also be answered by sophisticated methods, such as generalized estimating equations (GEE) or random coefficient analysis/multilevel analysis. The advantage of the sophisticated methods is that all available data is included in the analysis, while with (M)ANOVA for repeated measurements (and therefore also with (M)ANCOVA) only those subjects with complete data are included. Another important advantage of the sophisticated analyses is that they are regression techniques, from which the effect estimates (i.e., the magnitude of the effect of the intervention) and the corresponding confidence intervals can be derived easily.

The general idea behind all statistical techniques to analyze longitudinal data is that because of the dependency of observations within a subject a correction must be made for "subject". The problem, however, is that the variable "subject" is a categorical variable that must be represented by dummy variables. Suppose there are 200 subjects in a particular study. This means that 199 dummy variables are needed to correct for subject. Because this is practically impossible, the correction for "subject" has to be done in a different way and the different longitudinal techniques differ from each other in the way they perform that correction.

Within GEE, the correction for the dependency of observations is done by assuming (a priori) a certain "working" correlation structure for the repeated measurements of the outcome variable (Zeger and Liang, 1986; Liang and Zeger, 1986). Depending on the software package used to estimate the regression coefficients, different correlation structures are available. They basically vary form an "exchangeable" (or "compound symmetry") correlation structure, i.e., the correlations between subsequent measurements are assumed to be the same, irrespective of the length of the measurement interval, to an "unstructured correlation structure". In this structure no particular structure is assumed, which means that all possible correlations between repeated measurements have to be estimated.

In the literature it is assumed that GEE analysis is robust against a wrong choice for a correlation structure, i.e., it does not matter which correlation structure is chosen, the results of the longitudinal analysis will be more or less the same (Liang and Zeger, 1993; Twisk, 2004). However, when the results of analysis with different working correlation structures are compared to each other, the magnitude of the regression coefficients are different (Twisk, 2003). It is therefore important to realize which correlation structure should be chosen for the analysis. Although the unstructured working correlation structure is always the best, the simplicity of the correlation structure also has to be taken into account. The number of parameters (in this case correlation coefficients) which needs to be estimated differs for the various working correlation structures. In the example dataset with six repeated measurements, for instance, for an exchangeable structure only one correlation coefficient has to be estimated, while for the unstructured correlation structure, 15 correlation coefficients must be estimated. As a result, the power of the statistical analysis is influenced by the choice for a certain structure. The best option is therefore to choose the simplest structure which fits the data well. The first step in choosing a certain correlation structure can be to investigate the within-person correlation coefficients for the outcome variable. It should be kept in mind that when analyzing covariates, the correlation structure can change (i.e., the choice of the correlation structure should better be based conditionally on the covariates).

Random coefficient analysis (Laird and Ware, 1982) is also known as multilevel analysis (Goldstein, 2003; Twisk, 2006), hierarchical modeling, or mixed effects modeling. As has been mentioned before, the general idea behind all longitudinal statistical techniques is to correct for "subject" in an efficient way. Correcting for "subject" actually means that for all subjects in the longitudinal study, different intercepts are estimated. The basic principle behind the use of random coefficient analysis in longitudinal studies is that not all separate intercepts are estimated, but that (only one) variance of those intercepts is estimated, i.e., a random intercept. It is also possible that not only the intercept is different for each subject, but that also the development over time is different for each subject, in other words, there is an interaction between "subject" and time. In this situation the variance of the regression coefficients for time can be estimated, i.e., a random slope for time. In fact, these kind of individual interactions can be added to the regression model for all covariates. In a regular RCT, however, assuming a random slope for the intervention effect is not possible, because the intervention variable is time-independent (Twisk, 2006). When a certain subject is assigned to either the intervention or control group, that subject stays in that group along the intervention period. An exception is the crossover trial, in which the subject is his own control and the intervention variable is time-dependent. In this situation the intervention effect can be different for each subject and therefore a random slope for the intervention variable can be assumed. For random coefficient analysis, one has to choose which coefficients have to be assumed random. This choice is easier than the choice for a working correlation structure in GEE analysis. This is due to the fact that most standard software, which can be used for random coefficient analysis, provides a likelihood ratio chi-square test using −2 log likelihood values of each model, which can be used to evaluate different models.

Comparison between GEE analysis and random coefficient analysis

Both GEE and random coefficient analyses are highly suitable to analyze longitudinal experimental data, because in both methods a correction is made for the dependency of the observations within one subject. The question then arises: Which of the two methods is better? Unfortunately, no clear answer can be given. For continuous outcome variables, GEE analysis with an exchangeable correlation structure is the same as a random coefficient analysis with only a random intercept. The correction for the dependency of observations with an exchangeable "working correlation" structure is the same as allowing individuals to have random intercepts. When the dependency of observations is slightly more complicated, GEE analysis with a different correlation structure can be used or random coefficient analysis with additional random regression coefficients for other variables (e.g., time). Although random coefficient analysis is slightly more flexible, it should be realized that "regular" random coefficient analysis is limited by the fact that the random regression coefficients are assumed to be normally distributed, i.e., the variance of the intercepts is estimated by assuming a normal distribution.

Correcting for baseline values in sophisticated analyses?

It was already mentioned before that when baseline values differ between the intervention and control group it is (sometimes) necessary to correct for these differences, i.e., to correct for the phenomenon of regression to the mean. So, when more than one follow-up measurement is considered, this means that the whole longitudinal development of the outcome variable over time is corrected for the baseline value (see equation 8).

$$Y_{it} = \beta_0 + \beta_1 X_i + \beta_2 Y_{it0} + \ldots.. \quad (8)$$

where: Y_{it} = observations for subject i at time t; β_0 = intercept; β_1 = regression coefficient for X_i; X_i = intervention variable; β_2 = regression coefficient for observation at $t0$; and Y_{it0} = observation for subject i at time t_0.

The other possibility is to use a so-called autoregressive model in which the whole development of the outcome variable is not corrected for the baseline value, but each measurement of the outcome variable for the value of the outcome variable one time-point earlier [see equation (9)]

$$Y_{it} = \beta_0 + \beta_1 X + \beta_2 Y_{it-1} + \ldots.. \quad (9)$$

where: Y_{it} =observations for subject i at time t; β_0 = intercept; β_1 = regression coefficient for X_i; X_i = intervention variable; Y_{it-1} = observation for subject i at time t-1; and β_2 = regression coefficient for observation at *t*-1 (autoregression coefficient).

The idea underlying the autoregressive model is that the value of an outcome variable at each time-point is primarily influenced by the value of this variable one measurement earlier. To estimate the "real" influence of the intervention variable on the outcome variable, the model should therefore correct for the value of the outcome variable at time-point *t*-1. In fact, with an autoregressive model, the "corrected" changes between subsequent measurements are compared between the intervention and the control group.

Although the longitudinal analysis of covariance is mostly used, it is questionable whether or not this is correct. In fact, the correction for baseline overestimates the therapy effect. When there are, for instance, two follow-up measurements, the short-term effect is doubled in the estimation of the overall therapy effect. This situation is illustrated in Figure 18.3. As can be

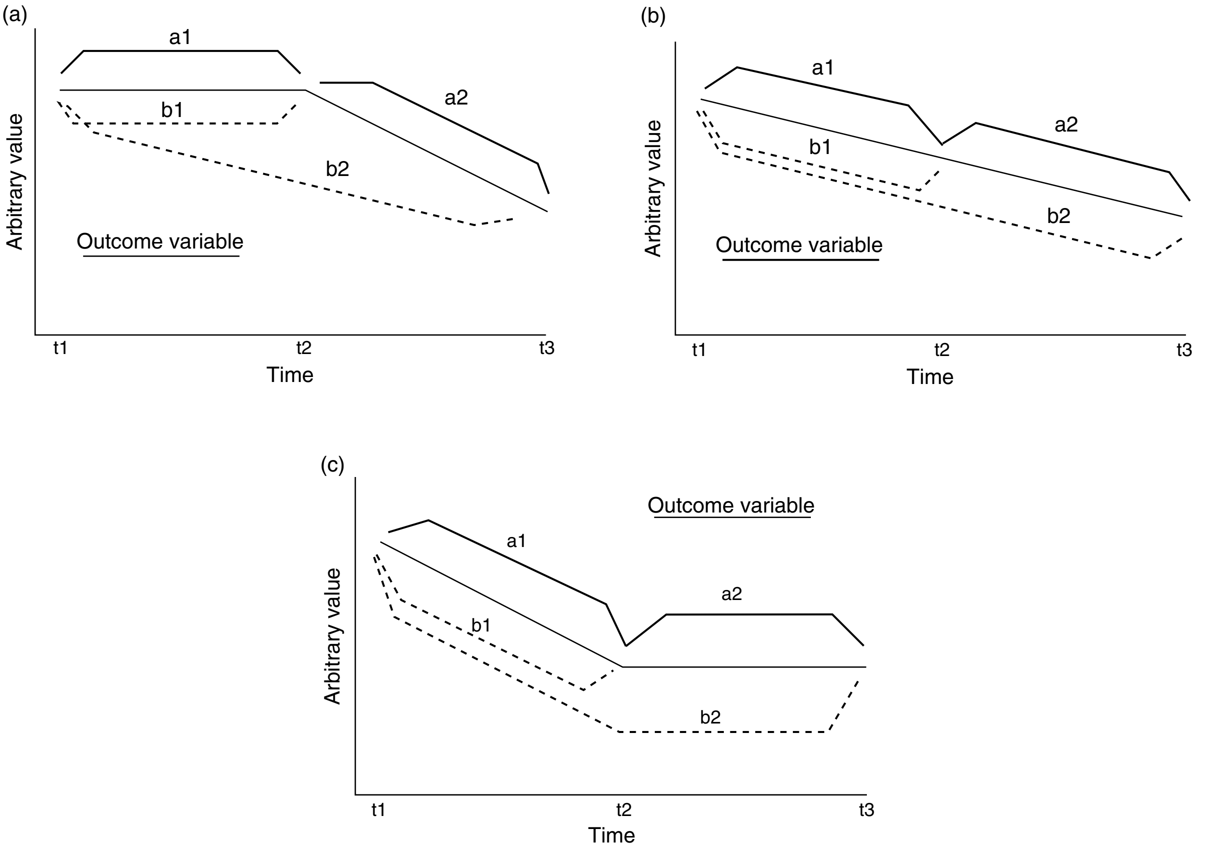


Figure 18.3 The difference between two approaches that can be used in the analysis of an experimental longitudinal study. The effects a1 and a2 are detected by autoregression analysis (equation 9), while the effects b1 and b2 are detected by the longitudinal analysis, correcting for the baseline value (equation 8). For Figure 18.3a, the two methods will show comparable results (a1= b1 and a2 = b2). For Figure 18.3b, the longitudinal analysis, correcting for baseline, will detect a stronger decline than the autoregression analysis (a1 = b1 and a2 < b2). The situation in Figure 18.3c will produce the same result as in Figure 18.3b (i.e., a1 = b1 and a2 < b2).

seen from this figure, this overestimation is not present when an autoregressive model is used.

3.2 Dichotomous outcome variables

In many experimental studies, the outcome variable of interest is not a continuous one, but a dichotomous one. A subject is recovered or not recovered, a subject experiences a relapse of a certain chronic disease or not, etc. When there is only one follow-up measurement, the analysis of a dichotomous outcome variable is not very complicated and can be performed with either logistic regression analysis or survival analysis. This has to do with the fact that mostly at baseline all subjects are the same; i.e., they are all patients with a certain disease the particular intervention is aiming at. The biggest problems occur when the dichotomous outcome variable can change over time, i.e., when the event of interest is recurrent. Figure 18.4 shows an example of a study in which the event of interest (i.e., treatment success) is recurrent.

Basically, the different statistical techniques to analyze recurrent event data can be divided into “naïve” techniques and longitudinal techniques. “Naïve” techniques are characterized by either ignoring the existence of recurrent events or ignoring the fact that the recurrent events within subjects are correlated. Longitudinal techniques on the other hand are characterized by the fact that the whole pattern of recurrent events over time is analyzed, taking into account that the recurrent events are correlated within subjects. Despite the fact that there are many statistical techniques available to analyze recurrent event data (Eisen, 1999), for most researchers it is rather difficult to choose the proper technique to answer the research question they are interested in. Reviewing the literature, it is rather surprising that most authors use “naïve” statistical techniques to analyze their study outcomes (Stürmer et al., 2000). The mostly used “naïve” statistical techniques are “naïve” in such a way that they do not use all available data, but only one observation for each patient. First, a logistic regression analysis can be performed in order to analyze the difference in the proportion of patients with “treatment success” at the end of the study. Second, a survival analysis (i.e., Cox proportional hazards regression) can be performed with the first experienced event (i.e., “treatment success”) and the time to that event as an outcome variable. A possible explanation for the popularity of using “naïve” techniques is that most longitudinal techniques are only described in specific statistical literature, which is difficult to understand for most (non-mathematical) researchers (Clayton, 1994; Lagakos, 1997). However, the general ideas behind these techniques are not as difficult as often suggested.

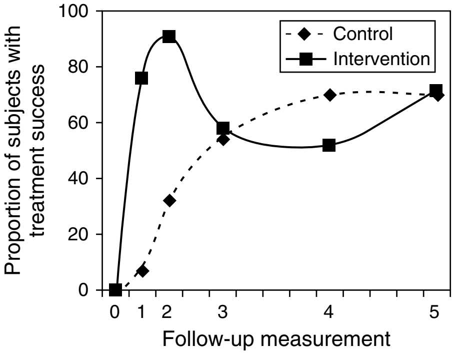


Figure 18.4 An experimental study with recurrent events

Longitudinal techniques

The longitudinal techniques to analyze recurrent events can be divided into survival approaches and (longitudinal) logistic regression approaches. Regarding survival approaches, Cox proportional hazards regression for recurrent events can be performed. Although there are different estimation procedures available (Kelly and Lim, 2003) the general idea behind Cox proportional hazards regression for recurrent events is that the different time periods are analyzed separately adjusted for the fact that the time periods within one patient are dependent. The idea of this adjustment is that the standard error of the regression coefficient of interest is increased proportional to the correlation of the observations within one patient. One of the problems using Cox proportional hazards regression for recurrent events is the question how to define the "time at risk". Especially when the events under study are not short-lasting events, but can be long-lasting, i.e., continuing over more than one time-point (i.e., they can be considered as "states"). Figure 18.5 shows a few possibilities to define the "time at risk" in a study where the dichotomous outcome variable is recurrent:: (1) the "counting approach"' Each time period is analyzed separately assuming that all subjects are at risk at the beginning of each period, irrespective of the situation at the end of the foregoing period; (2) the "total time" approach. This is comparable to the "counting" approach. However, in the "total time" approach, the starting point for each period is the beginning of the study; (3) the "time to event" approach. For instance, when the event of interest is treatment success, in this approach only the transitions from "no treatment success" to "treatment success" are taken into account. So, if for a subject (or patient) the treatment was "successful" at the first follow-up and stays "successful" at all repeated measurements, only the first time is taken into account in the analysis. When for another subject (or patient) the treatment was "successful" at the first follow-up, and "not successful" at the second follow-up, that particular subject (or patient) is again at risk from the first follow-up onwards until the treatment for that subject (or patient) is "successful" for the second time, or until the follow-up period ended.

Regarding the logistic regression approaches, the two techniques that have already been discussed for continuous outcome variables can also be used to analyze the recurrent event data, i.e., GEE analysis and random coefficient analysis (see page 284). When dichotomous outcome variables are considered, a logistic version

(a)

ID	time6	time12	time24	time36	time52
1	0	1	0	1	1
..					

'*total time approach*'

ID	event	time
1	0	6
1	1	12
1	0	24
1	1	36
1	1	52

(b)

ID	time6	time12	time24	time36	time52
1	0	1	0	1	1
..					

'*counting approach*'

ID	event	time
1	0	6
1	1	6
1	0	12
1	1	12
1	1	16

(c)

ID	time6	time12	time24	time36	time52
1	0	1	0	1	1
..					

'*time to event approach*'

ID	event	time
1	–	–
1	1	6
1	–	–
1	1	24
1	–	–

Figure 18.5 Possible definitions of the time at risk to be analyzed with Cox regression for recurrent events for a subject in a hypothetical study who experienced an event at time12, relapsed at time24, experienced an event again at time36 which continued till the end of the study

of both GEE analysis and random coefficient analysis is available. It should be noted that for dichotomous outcome variables, (M)ANOVA for repeated measurements can not be used. In contrast to the analysis with a continuous outcome variable, the two longitudinal logistic regression techniques (i.e., GEE analysis and random coefficient analysis) lead to different results. Basically, both "longitudinal" techniques take all measurements into account, and use a logistic regression approach with a correction for the dependency of the observations. Again either by assuming a certain "working" correlation structure (GEE analysis) or by allowing random regression coefficients (random coefficient analysis). Because a logistic regression approach is used, the regression coefficients derived from both techniques can be transformed into odds ratios. The difference between the two techniques is that GEE analysis is a so-called population average approach, while random coefficient analysis is a so-called subject specific approach (Twisk, 2003). The different estimation procedures cause the difference in the magnitude of the odds ratios, which is always in favor of the random coefficient analysis, i.e., the "effects" estimated with random coefficient analysis are always bigger than the "effects" estimated with GEE analysis. (see Figure 18.6). Because the standard errors

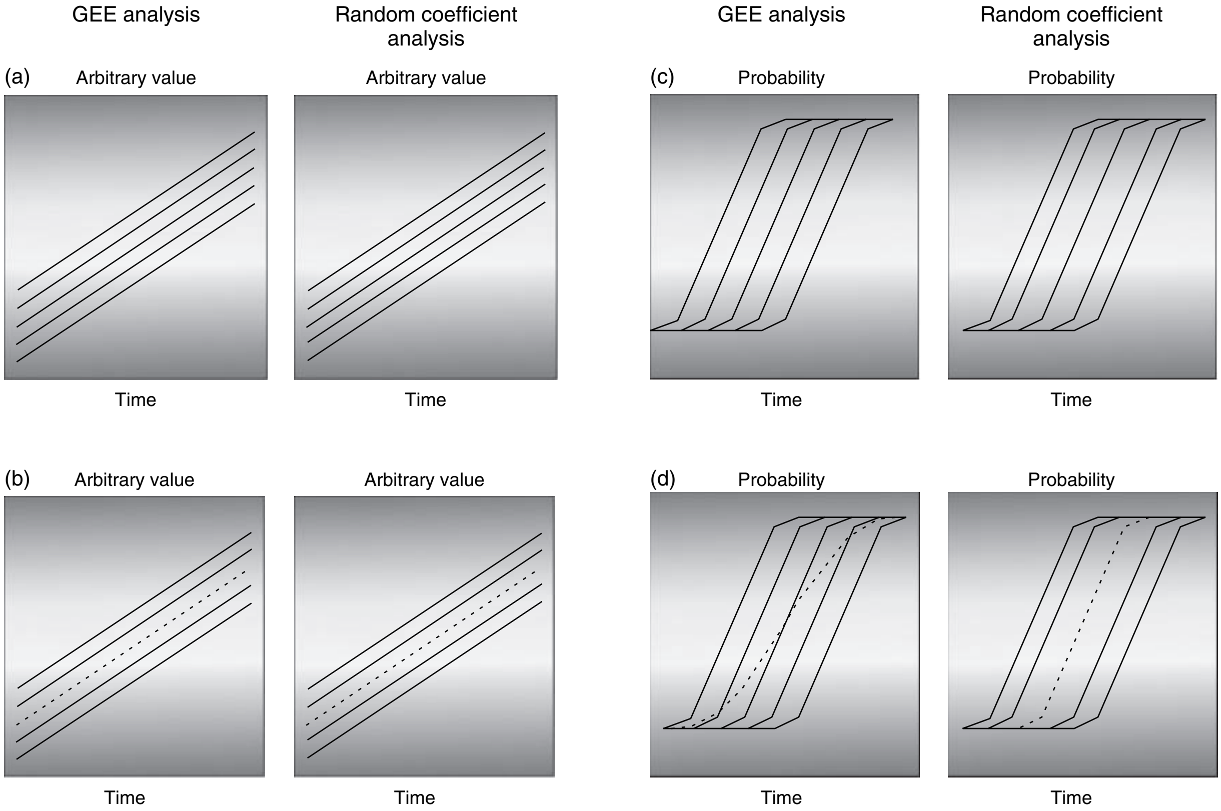


Figure 18.6 The "population average" approach of GEE analysis and the "subject specific" approach of random coefficient analysis, illustrating both the situation with a continuous outcome variable (Figure 18.6a and b), and the situation with a dichotomous outcome variable (Figure 18.6c and d).

are also bigger for the random coefficient analysis (and therefore the 95% confidence intervals wider), the corresponding p-values are not much different, and when conclusions are based on these p-values, they will not be much different when using GEE analysis or random coefficient analysis. However, when the conclusions are based on the magnitude of the odds ratios, the conclusions will differ remarkably between the two techniques.

It should further be noted that the estimations of the regression coefficients (i.e., odds ratios) with random coefficient analyses of recurrent events can be very complicated and often lead to unstable results. Furthermore, the results of these analyses can differ between software packages (Yang and Goldstein, 2000; Lesaffre and Spiessens, 2001; Twisk, 2003).

Comments

In the previous section an overview was given of different techniques that can be used for the analysis of recurrent event data. The simplest and probably most illustrative way of describing recurrent event data is plotting the proportion of subjects experiencing the event at each time-point (see Figure 18.4) or showing the different response patterns observed. Although both can give a nice overview, it is difficult to analyze these patterns. So, therefore, several statistical analyses can be performed on the recurrent event data. It is striking that the techniques that use all available data can give totally different results than the techniques that use only part of the data (Twisk et al., 2005). The choice for a particular technique highly depends on the research question. With the logistic regression analysis using the data assessed at the end of the follow-up period the long-term effect of the intervention is analyzed, while with the "naïve" Cox regression analysis the short-term effect of the intervention is analyzed. With the longitudinal techniques that use all available data, the overall intervention effect, i.e., the whole development over time, is analyzed. Note, too, that Cox regression analysis outputs hazard ratios, which are different from and cannot be used interchangeably with the odds ratios output in logistic regression analysis.

A major problem of using Cox regression for recurrent events on the other hand is the assumption of proportional hazards over time; an assumption that does not hold in many situations. When the proportional hazards assumption does not hold, it is possible to divide the follow-up period into several subperiods and calculate different hazard ratios for each subperiod. Furthermore, compared to the longitudinal logistic regression approaches the possibilities to correct for the dependency of observations in using Cox regression are rather limited. In fact the correction only influences the standard error of the regression coefficient, i.e., the width of the 95% confidence interval around the hazard ratio. The point estimate is equal to the point estimate derived from an analysis when the observations are considered to be independent.

All statistical techniques discussed in light of the analysis of recurrent events are either an extension of Cox proportional hazards regression or logistic regression. Therefore, issues such as effect modification and confounding can be handled in exactly the same way as in the "classical" application of these techniques. Of course, due to the longitudinal nature of the data, possible effect modifiers or confounders can be time-independent as well as time-dependent.

The example shown in Figure 18.4 is an example of a study with single type event data. Only one kind of event (i.e., "treatment success") is used as outcome. Although the interpretation of the results is slightly different, it is obvious that the same kind of approaches as described in this paper can be used for analysis of multitype event data such as tumors at different sites, different kinds of infection, etc. (Wei and Glidden, 1997).

4 General recommendation

The way the results of an experimental study should be analyzed highly depends on the research question of interest. If one is only interested in a particular short-term or long-term result, simple techniques are highly appropriate. However, if the development of a particular outcome is of interest, the longitudinal statistical techniques are necessary to answer the accompanying research question. For continuous outcome variables, the results of GEE analysis and random coefficient analysis are comparable, but for dichotomous outcome variables, this is not the case. When a dichotomous outcome variable is analyzed and when the longitudinal dataset consists of discrete time-points, GEE analysis or random coefficient analysis can be used, but GEE is to be recommended because of the availability of the population average approach and the relatively "simple" and robust estimation procedures compared to random coefficient analysis. When the events can occur continuously, Cox proportional hazards regression for recurrent events must be used, but special attention has to be given to the definition of the "time at risk" and to the assumption of proportional hazards.

References

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychology Bulletin*, 105: 290–301.

Agresti, A. (1999). Modeling ordered categorical data: Recent advantages and future challenges. *Statistics in Medicine*, 18: 2192–2207.

Bereiter, C. (1963). Some persistent dilemmas in the measurement of change. In C. W. Harris (ed.), *Problems in the Measurement of Change*, pp. 3–20. Madison: University of Wisconsin Press.

Clayton, D. (1994). Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research*, 3: 244–262.

Cronbach, L. J. and Furby, L. (1970). How should we measure "change"—or should we? *Psychology Bulletin*, 74: 68–80.

Eisen, E. A. (1999). Methodology for analyzing episodic events. *Scandinavian Journal of Work Environment, and Health*, 25 Suppl. 4: 36–42.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd edn. London: Edward Arnold.

Gottman, J. M. (1995). *The Analysis of Change*. Mahwah, NJ: Lawrence Erlbaum.

Kelly, P. J. and Lim, L-Y. (2003). Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine,* 19: 13–33.

Lagakos, S. (ed.) (1997). Statistical methods for multiple events data in clinical trials. *Statistics in Medicine*, 16: 831–964.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38: 963–974.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, 50: 325–335.

Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73: 45–51.

Liang, K-Y. and Zeger, S. L. (1993). Regression analysis for correlated data. *Annual Review of Public Health*, 14: 43–68.

MacKnight, C. and Rockwood, K. (2000). Rasch analysis of hierarchical assessment of balance and mobility (HABAM). *Journal of Clinical Epidemiology*, 53: 1242–1247.

Plewis, I. (1985). *Analyzing Change: Measurement and Explanation Using Longitudinal Data*. Chichester, UK: Wiley.

Raczek, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., Apolone, G., Bech, P., Brazier, J.E., Bullinger, M. and Sullivan, M. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA Project International Quality of Life Assessment. *Journal of Clinical Epidemiology*, 51: 1203–1214.

Rogosaa, D. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J.M. Gottman (ed.), *The Analysis of Change*, pp. 3–66. Mahwah, NJ: Lawrence Erlbaum.

Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology*. Philadelphia: Lippincott-Raven.

Schnell, D. J., Magee, E. and Sheridan, J. R. (1995). A regression method for analyzing ordinal data from intervention trials. *Statistics in Medicine*, 14: 1177–1189.

Sonn, U. and Svensson, E. (1997). Measures of individual and group changes in ordered categorical data: Application to the ADL staircase. *Scandinavian Journal of Rehabilitation Medicine*, 29: 233–242.

Stucki, G., Daltroy, L., Katz, J. N., Johannesson, M. and Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology*, 49: 711–717.

Stürmer, T., Glynn, R. J., Kliebsch, U. and Brenner, H. (2000). Analytic strategies for recurrent events in epidemiologic studies: Background and application to hospitalization risk in the elderly. *Journal of Clinical Epidemiology*, 53: 57–64.

Svensson, E. (1998). Ordinal invariant measures for individual and group changes in ordered categorical data. *Statistics in Medicine*, 17: 2923–2936.

Twisk, J. W. R. (2003). *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge, UK: Cambridge University Press.

Twisk, J. W. R. (2004). Longitudinal data analysis: A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology*, 19: 769–776.

Twisk, J. W. R. (2006). *Applied Multilevel Analysis*. Cambridge, UK: Cambridge University Press.

Twisk, J. W. R and Proper, K. (2004). Evaluation of the results of a randomized controlled trial: How to define changes between baseline and follow-up. *Journal of Clinical Epidemiology*, 57: 223–228.

Twisk, J. W. R., Smidt, N. and Vente, W. de. (2005). Applied analysis of recurrent events: A practical overview. *Journal of Epidemiology and Community Health*, 59: 706–710.

Wei, L. J. and Glidden, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine*, 16: 833–839.

Yang, M. and Goldstein, H. (2000). Multilevel models for repeated binary outcomes: Attitudes and voting over the electoral cycle. *Journal of the Royal Statistical Society*, 163: 49–62.

Zeger, S. L. and Liang, K-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42: 121–130.

This page intentionally left blank

Part IV

Description and Measurement of Qualitative Change

This page intentionally left blank

Chapter 19

Analyzing longitudinal qualitative observational data

Johnny Saldaña

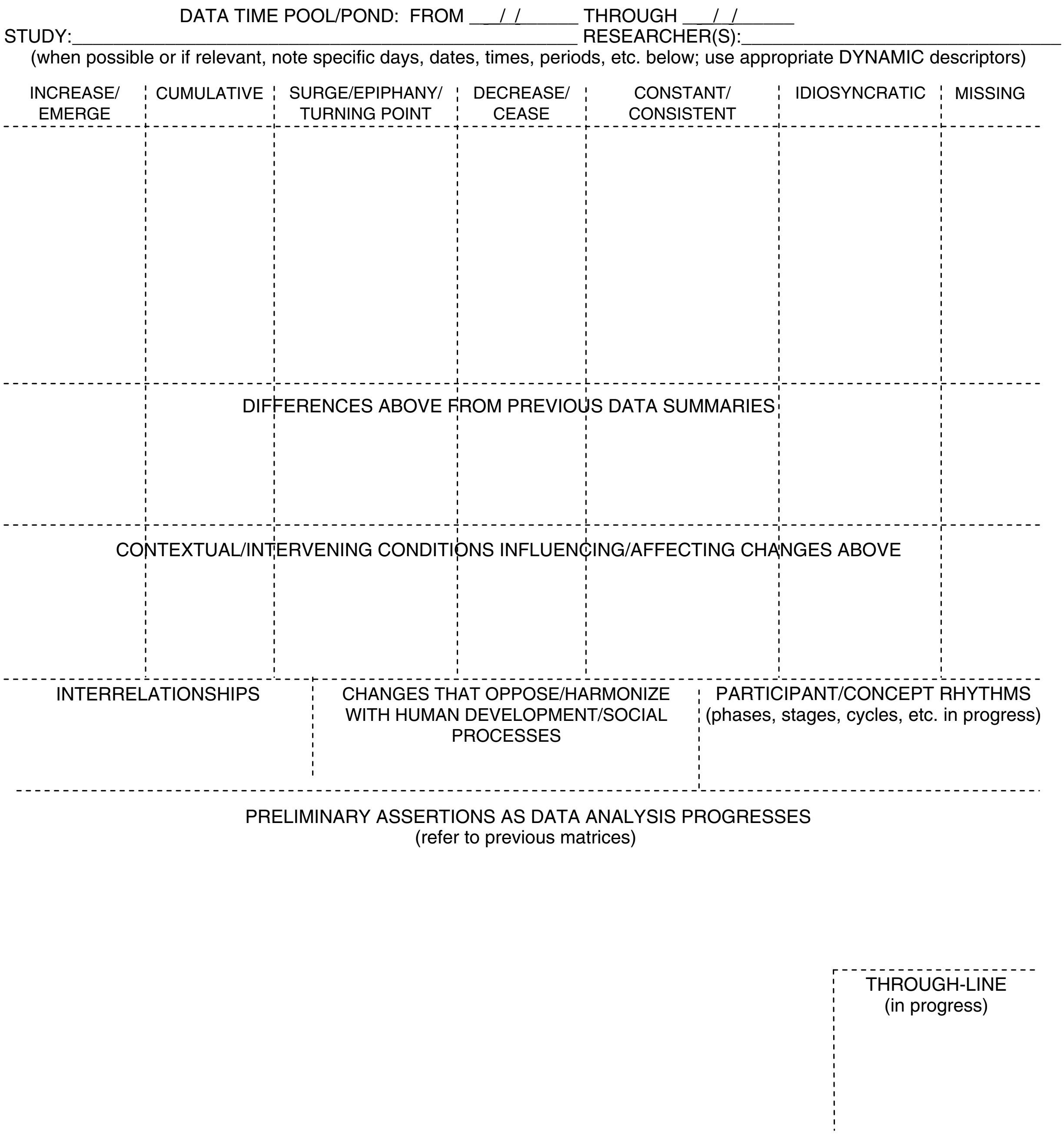
1 Analyzing longitudinal qualitative observational data

This chapter assumes that readers already have experience writing detailed and useful fieldnotes through a longitudinal time period (see Kathleen M. DeWalt and Billie R. DeWalt's *Participant Observation: A Guide for Fieldworkers* [2002]; Robert M. Emerson, Rachel I. Fretz, and Linda L. Shaw's *Writing Ethnographic Fieldnotes* [1995]; and James P. Spradley's *Participant Observation* [1980] for examples). What I address below is my signature approach to analyzing fieldnote data derived from long-term observations of participants in social settings. Due to restrictions of chapter length, I cannot detail the theoretical underpinnings and nuances of my eclectic methods of longitudinal qualitative data analysis. Instead, I will focus on pragmatic matters and refer readers to my book, *Longitudinal Qualitative Research: Analyzing Change through Time* (2003) for a more thorough discussion of the complexities involved with this topic.

1.1 Change processes

I propose that researchers are more likely to discern participant or environmental change (if any) from longitudinal qualitative data if we code them using indicators of change processes. These processes were adapted from a literature review of quantitative and qualitative longitudinal methods (e.g., Fullan, 2001; Huber and Van de Ven, 1995; Kelly and McGrath, 1988; Royce and Kemper, 2002; Menard, 1991; Miles and Huberman, 1994; Strauss and Corbin, 1998; Sztompka, 1993), and constructed from personal longitudinal research experiences (Saldaña, 1995, 1996, 1997, 1998a, 1998b, 2005). Just as there is statistical increase, decrease, constancy, idiosyncrasy and the like in quantitative data, so too can there be qualitative increase, decrease, constancy, idiosyncrasy and the like within and among participants in social settings.

Coding qualitative data is a preliminary form yet vital function of analysis, for the process involves the researcher's comparison of data from apportioned time periods and attributing difference and thus possible change (if any) to written field observations. In the methods I propose in this chapter, coding and consequent analysis adhere to a particular repertory of change processes, which can be seen in abbreviated form in Figure 19.1—a longitudinal qualitative data summary matrix. This matrix enables researchers to pool, summarize, and transfer qualitative fieldnote data from a



DATA TIME POOL/POND: FROM ___/_/______ THROUGH ___/_/______

STUDY:____ RESEARCHER(S):______________________________

(when possible or if relevant, note specific days, dates, times, periods, etc. below; use appropriate DYNAMIC descriptors)

INCREASE/ EMERGE	CUMULATIVE	SURGE/EPIPHANY/ TURNING POINT	DECREASE/ CEASE	CONSTANT/ CONSISTENT	IDIOSYNCRATIC	MISSING

DIFFERENCES ABOVE FROM PREVIOUS DATA SUMMARIES

CONTEXTUAL/INTERVENING CONDITIONS INFLUENCING/AFFECTING CHANGES ABOVE

INTERRELATIONSHIPS	CHANGES THAT OPPOSE/HARMONIZE WITH HUMAN DEVELOPMENT/SOCIAL PROCESSES	PARTICIPANT/CONCEPT RHYTHMS (phases, stages, cycles, etc. in progress)

PRELIMINARY ASSERTIONS AS DATA ANALYSIS PROGRESSES
(refer to previous matrices)

THROUGH-LINE
(in progress)

Figure 19.1 Longitudinal qualitative data summary matrix

selected time period of a longitudinal study onto a readable one-page format.

Imagine that one matrix page holds summary observations from three months' worth of fieldwork. And if the study progresses through two years, then there would be eight pages total of longitudinal qualitative data. I've developed several metaphors to illustrate how multiple matrix sheets arranged in chronological order work together as a strategy for analyzing participant change. Think of each three-month page as an animator's cartoon cell, whose artwork changes subtly or overtly with each successive drawing to suggest movement and change. Or, imagine that each matrix sheet is a monthly page from a calendar, which suggests a chronological progression of time and change as each page is turned. Or, imagine that each matrix page is a photograph of the same child taken at different intervals across time, so that each successive photo reveals growth and development. Finally, consider that each matrix might also represent a page from a personal diary, whose chronological flow of entries tells a narrative of what happened and then what happened next.

The matrix may, at first glance, appear conceptually or theoretically abhorrent to some, and admittedly there is a "breaking in" period to using it effectively. Nevertheless, it's a model I've tested and modified from its original design (Saldaña, 2003, pp. 54–55, 172; Saldaña, 2005), and now propose it as a system for longitudinal qualitative data entry. Note that I did not write that this is a system for longitudinal qualitative data *analysis*, for the analytic thinking must still be done by you. The primary functions of the matrix are data management and to reduce and categorize field observations from a selected time period to assist the researcher with "analysis at a glance," if you will.

1.2 The matrix

The chapter will now focus on the inventory of change processes from Figure 19.1 in more detail. Each cell also provides an opportunity to briefly discuss major concepts related to longitudinal qualitative data analysis.

DATA TIME POOL/POND

The top entry records the start and end dates of a particular period of time from the longitudinal research study. (A "pond" is a smaller or shorter portion from a "pool" of data.) Each pool/pond and thus page can consist of a month's, year's, or even several years' worth of summary data. Though this may be stating the obvious, it is critical that you routinely note the date and time of all observations in your fieldnotes with attention to such time-related contexts as the day of the week, the number of days into a school year, the season, holiday or ceremonial preparations in progress, ages of participants (in years and months), and culturally specific conceptions of time (see Levine, 1997).

Each matrix page does not have to include data from a standardized timeframe or regularly fixed time interval, such as the end of one academic year or grade level for studies in education. Depending on the length of your project and as you're observing and collecting data, you may notice clusters of similarities within particular varying timeframes such as three months, seven months, two years, then ten months. This researcher-generated division of total fieldwork time is a task that "enables you to examine the dynamics of change such as duration, frequency, and tempo, which then supports the development of conceptual phases, stages, cycles or other rhythms of human action" (Saldaña, 2003, p. 160).

The first matrix in the series—the "genesis" page—is the most difficult to complete, for it contains baseline data for comparison with other future matrix pages. The researcher relies on fieldnotes that have accumulated through the first phase, stage, or cycle of fieldwork that suggest change. For example, one of my longitudinal case studies focused on the artistic development of Barry (pseudonym), a male from ages 5 through 26 (see Saldaña, 1995, 1998a, 1998b, 2005). Though I had been analyzing the data periodically at various time periods during this young man's life course, the final analytic venture happened when he turned age 26, the time when both Barry and I agreed that the study had come to a conclusion. The first pool of data for the genesis matrix page consisted of observations from ages 5.5 through 12.5 [years.months]; the second pool from ages 12.6 through 16.4; the third pool from ages 16.5 through 18.5; the fourth pool from ages 18.5 through 23.3; the fifth pool from ages 23.4 through 23.11; and the sixth and final pool from ages 24 through 26.1. The rationale for the divisions were as follows:

- *Ages 5.5–12.5*: Barry's elementary school years, grades K–6, consisted of data that tracked his participation in and responses to theatre. Though each grade level was a pond of data for comparison with other grade levels, the total grades K–6 pool was selected for one matrix page since this seemed a developmentally-appropriate division as an educational study.
- *Ages 12.6–16.4*: There would be no direct observation of the participant for the next four years, yet data from these stages of Barry's life course were provided retrospectively in later interviews. This second pool included two key epiphanies and thus merited a separate matrix page.
- *Ages 16.5–18.5*: Like the first data pool, the total grades 10–12 pool—Barry's high school years—was selected for one matrix page since this seemed a developmentally-appropriate division as an educational study.
- *Ages 18.5–23.3*: This fourth pool of data consists of what I labeled, in retrospect, a "searching" phase within Barry's life course. The time period covers a few years of post-secondary school employment and community college education, but the end of the time period was determined by a key interview with the participant—a data gathering experience that suggested he was on the verge of discovery within his life course.
- *Ages 23.4–23.11*: The fifth pool of data was determined solely by the interview schedule—the time period between Barry's last interview at age 23.3, and the next one at age 23.11. There was no direct participant observation during this time.
- *Ages 24–26.1*: The final pool of data consists of a time period in which Barry's life and career goals crystallized. It includes both university education and an epiphanic moment of discovery, but its conclusion at 26.1 years of age was determined solely by an agreement between the participant and myself that the formal study had come to a satisfactory conclusion.

Hence, the division of data time pools/ponds for matrix pages can consist of such traditional and socially constructed periods as: schooling, periods between data gathering opportunities, epiphanies/turning points in a participant's life course, and retrospective division of a participant's life course into phases, stages, or cycles of human development.

Most longitudinal studies with children and adolescents report any observed changes according to their ages or grade levels—traditional developmental markers that provide some sense of standardization across several disciplines such as psychology, education, and sociology. Recent methodology (e.g., Levine, 1997; Tudge and Hogan, 2005, p. 114), however, notes that varying cultural backgrounds of participants, ranging from ethnicity to nationality, may influence and affect change across time and thus present contradictory findings when we compare individuals or groups. Likewise, CONTEXTUAL AND INTERVENING CONDITIONS (see discussion below) such as nontraditional home environments or personal family tragedy, may present different change patterns that do not conform to theoretically universal patterns of child development.

I recommend that you accumulate a minimum of three matrix pages of data for analysis to support any observations and assertions of change. Two matrix pages provide only a "then and now" comparison, which potentially glosses over the processes of participant change. Three or more pages enable you to chronologically path the journey of participant change (if any) through time. Change does not always happen during regularly fixed time intervals, and change can be such a slippery and elusive process to track that it can only be noticed in retrospective data analysis rather than during the moment or period it occurs. It is the researcher's intuitive but strategic choice to decide which data from the corpus constitute a pool or pond.

STUDY and RESEARCHER(S)

Inserting the names of the longitudinal study and its researcher(s) is fairly obvious, but let me offer a few notes on research team members jointly analyzing longitudinal qualitative observational data.

Colleagues, whether fellow researchers or the participants themselves, are excellent resources for deep and insightful conversations about the data. This matrix and its particular change processes lend themselves to productive dialog about the actions and phenomena observed through time. Reaching consensus on whether an observation of change is an INCREASE or DECREASE (see discussion below) helps focus the interpretations. But be aware that multiple researchers responsible for analyzing longitudinal qualitative data can turn into a longitudinal enterprise in and of itself if the dialog becomes stalled, particularly during coding processes. All research team members should monitor themselves and each other on the progress of their analytic endeavors and ensure that everyone is working toward the ultimate goal: What, if anything, has changed through time?

DYNAMIC Descriptors

Dynamics refer to the dimensions and variability of qualitative data. This note at the top of the matrix serves as a reminder to choose descriptive words carefully as entries are made in each cell.

The qualitative researcher's observations of change are highly interpretive acts. Few of us would disagree that the destruction of the US World Trade Center buildings on September 11, 2001, or the devastation inflicted by hurricane Katrina on the city and people of New Orleans in late August 2005, were events of tremendous magnitude and thus classified as EPIPHANIES or TURNING POINTS using these typical categories of change. But my word choices of "destruction," "devastation" and "tremendous magnitude" are, in fact, value judgements of my observations, for I could have also used such terms as "leveling," "annihilation," and "catastrophic proportions"—synonymous descriptors yet evocative of different meanings.

Just as the dynamics of observation and change are highly interpretive acts, so too are the interpretations of whether particular observations are increases, decreases, or consistencies in certain actions. For example, one researcher may note that an adolescent's clothing—from goth one year to punk the next—has become "*increasingly* radical," while a second may note that the wardrobe has become "*decreasingly* mainstream," while yet a third observes that the clothing styles have remained "*consistently* nontraditional."

Quantitative research has the advantage of scaling change on continua such as "none at all" to "somewhat" to "very much." Qualitative research can also apply such descriptors to change processes, but since the paradigm's advantage is language rather than statistics, words become powerful and, arguably, more accurate indicators of observed change. When we say that a person has become "more conservative" through time, this implies we were focusing narrowly on his conservatism to begin with and recorded change as a simple increase, when we could have plotted a personal change of ethos using a qualitative continuum: from "hardcore ACLU-er" to a "self-proclaimed 'independent'" to a "battle-weary liberal" to "socially conscious yet politically moderate." Through dynamic descriptors, we extend beyond a "then and now" model and illustrate not just incremental participant change but participant transformation. And if we are concerned in our analysis with such time orientations as tempo, frequency, or duration of actions, there is a difference, for example, between describing a participant in an administrative position who provides "two additional yet difficult years of service," and one who provides "twenty-four grueling months of self-sacrificing service above and beyond the call of duty that deteriorated his physical and emotional health."

I have found that the search for just the right word can sometimes confound my analysis and writing, but it forces me to think deeply about how I am representing and presenting my observations of change. This reflection is both an introspective audit—a "reality check"—of qualitative data, and an exercise in confronting one's credibility and trustworthiness with language choices.

The discussion now turns to the data entry cells themselves—bullet points of observations that summarize and capture the essences and essentials of participant change (if any) through time.

INCREASE/EMERGE

This cell includes summary observations that answer the question, *What increases or emerges through time?* This may include both quantitative and qualitative observations, yet primarily the latter. Increases in an individual's age and weight are examples of quantitative change, but related qualitative increases and emergences may include such participant observations as "difficulty moving," "worn and tired facial expressions," and "darker wardrobe colors" for the particular data pool. Richer inferential meanings from these patterns can be constructed by asserting more than just "getting older."

As an extended example, in the longitudinal qualitative study with Barry, the first data pool at ages 5.5–12.5 included the following as increases and emergences:

- additional theatre-viewing experiences beyond the treatment
- parental involvement in nurturing his theatre interest
- at age 12, reflecting on career choices (actor, writer, "think tank")
- at ages 11.8–12.5, victim of bullying by peers
- at age 12, counseling for withdrawal and depression.

Barry was not formally tracked from ages 12.6 through 16.4, since the initial longitudinal study was completed. In retrospective accounts during later years, he recalled two key epiphanies (a suicide attempt and his first formal performance experience) during the period I did not directly observe him. Related INCREASE/EMERGE data from the second pool included:

- at ages 12.6–14.6, anxiety from peer bullying
- hair length
- new: smoking, illegal drug use
- age 14.7, attitude "renaissance" from first and future performance opportunities.

Follow-up and direct participant observation was initiated during Barry's final secondary school years. The third data pool at ages 16.5–18.5 listed the following increases and emergences:

- roles in theatre productions
- concentration during performance work
- "passion" for the art form
- new: mentorship from theatre teachers
- leadership skills
- new: questioning his spiritual faith/belief systems.

The fourth data pool at ages 18.5–23.3 included the following as increases and emergences:

- searching for "artful living"
- service as a summer camp counselor for special populations
- attending a different church but same faith
- attending community college for general studies
- learning ASL [American Sign Language]
- exploring drama therapy as a career
- new: personal credit card
- new: prescription medication for bipolarity.

The fifth data pool at ages 23.4–23.11 listed the following as his increases and emergences in actions:

- deciding between social work and urban sociology as possible majors at the university
- providing urban ministry for youth
- new: eyebrow piercing, facial hair, spiked hair style.

The sixth and final pool of data at ages 24–26.1 listed as his increases and emergences:

- university education: pursuing a bachelor's degree in social work with a minor in religious studies
- new: tattoo on left arm—"fight, race faith" (from 2 Timothy 4:7)
- preaching occasionally at Sunday worship services
- new: disclosure of his father's past spiritual abuse
- working for "social justice"
- new: rock climbing as a hobby.

An INCREASE as change differs from the next two cells—observations that are CUMULATIVE, and observations that are SURGES/EPIPHANIES/TURNING POINTS. An increase or emergence in qualitative change is a phenomenon or participant action that appears or transforms in subtle, smooth, or expected ways. Cumulative change is a transformation in participant quality as a result of *successive* experiences through time. Surges, epiphanies, and turning points suggest change that is uncharacteristically rapid or sudden, a revelatory or insightful experience for the participant, or an event of such significant magnitude that it redirects the course and flow of a participant's consequent actions.

CUMULATIVE

This cell includes summary observations that answer the question, *What is cumulative through time?* Examples may include: a fifth-year teacher's finely developed expertise with classroom management based on her four years of previous instructional experience; a child's oral language fluency by age 12; and Barry's accumulated coursework experiences from community college and university education.

As noted above, cumulative affects are changes as a result of *successive* experiences through time. A minimum of three matrix pages permits you to track how selected common phenomena or actions transform in quality. This three-cell "time triangulation" (Saldaña, 2003, p. 164) provides a data-trail of support for any assertions related to this type of change process. Cumulative change, however, is not always a smooth path, particularly with human development's and social life's unexpected twists and turns along the journey. Hence, the next set of change processes needs to be taken into consideration during analysis.

SURGE/EPIPHANY/TURNING POINT

This cell includes summary observations that answer the question, *What kinds of surges, epiphanies, or turning points occur through time?* This category of change includes events, experiences, or personal revelations of *magnitude* in a participant's life course which significantly alter attitude, value, and belief systems;

initiate future actions in different directions; or lead to exponential and multiple consequent changes. A surge may be interpreted as an accelerated increase, decrease, etc.; an epiphany as an impacting event or revelatory moment; and a turning point as a juncture in the life course that separates significantly different courses of action. Examples may include: the devastation in the city of New Orleans after hurricane Katrina; a woman's new life course directions after a divorce; or Barry's suicide attempts during his junior high and post-secondary school years.

DECREASE/CEASE

This cell includes summary observations that answer the question, *What decreases or ceases through time?* Examples may include: a decline in faculty collegiality after disagreement over new directions for a degree program; a decrease in self-motivation at the workplace; or Barry's decrease and eventual cessation of illegal drug use.

A cautionary note: since written and recorded data themselves serve as evidence for our assertions, keep in mind that fieldnotes through time, particularly in the latter phases, stages, or cycles of fieldwork, do not always record what was gathered at the onset. Hence, what may seem "less of" something as time continues may actually be due to the fact that the fieldnotes themselves simply have not maintained a continuous record of initial phenomena or actions.

CONSTANT/CONSISTENT

This cell includes summary observations that answer the question, *What remains constant or consistent through time?* Note that it is the largest of the descriptive data cells in the top row, for my own research experiences suggest that a sizable portion of fieldnote observations may remain constant and consistent through time (assuming no actions of consequential magnitude, such as epiphanies). The "recurring and often regularized features of everyday life" (Lofland et al., 2006, p. 123) are, after all, what provide "analytic significance" for the social scientist. Examples may include: routine workplace methods prescribed by a standard operations manual and followed by employees; a continuous, perceived "lack of time" by a working mother; and Barry's residence with his parents from birth through age 26.1.

A culture is not a fixed system; it is an ever-changing and ever-evolving social organization. Yet constancy and consistency suggest the existence and thus documentation of predictable patterns in a social setting, and change may or may not occur within the established routines and rituals of daily life. Interpretation of this stability can range from the secure and steady to the rigid and stagnant, revealing "either something significant at work or nothing out of the ordinary" (Saldaña, 2003, p. 165). Nevertheless, we cannot discern what is changing unless we also know what is not changing. And when something does indeed change, we need to document how selected areas of participant constancy and consistency may have been influenced and affected.

IDIOSYNCRATIC

This cell includes summary observations that answer the question, *What is idiosyncratic through time?* Anomalies occur in the daily routine lives of participants, and these deviations merit their own cell for data entry. These do not include phenomena or actions of magnitude such as epiphanies, but rather the "inconsistent, ever-shifting, multidirectional and, during fieldwork, unpredictable" (Saldaña, 2003, p. 166). Examples may include: a child's sudden yet fleeting interest in a fictional superhero; a temporary relocation of personnel and operations while an office facility undergoes repairs; and Barry's irregular mood swings while under prescription medication for bipolarity.

As noted above, our life journeys do not follow a smooth path. Irregularities and uneven slopes of development are "givens" in the human condition, and thus the idiosyncratic deservedly earns its place in the matrix. The idiosyncratic is not a deviation from a smooth observable pattern of change—the idiosyncratic *is* part of the pattern if its occasional or even frequent occurrence can be explained.

MISSING

This cell includes summary observations that answer the question, *What is missing through time?* During participant observation, the researcher certainly notes what is present in the social setting, but should also consider what is absent or missing. Examples may include: a lack of material resources for more effective classroom instruction; a vacant staff position causing other employees to carry the workload; or Barry's lack of theatre experiences during his junior high school years. "Just because something is missing doesn't mean that nothing is being influenced and affected" (Saldaña, 2003, p. 166). Certainly we can identify a multitude of absent phenomena and actions in any social setting, but in this cell we note that which is most possibly and plausibly missing as they relate to what is present.

Once a time-apportioned matrix page of summary longitudinal qualitative data have been entered in the first row, the next step is to compare (with the exception of the first or genesis page) how these current observations differ from those logged in other matrix pages about the participants' or environment's chronological past (and future).

DIFFERENCES ABOVE FROM PREVIOUS DATA SUMMARIES

For each cell in the top portion of the matrix, similar observations from previous matrices are used to compare the current conditions. This row spirals the researcher back to previous observations of similar change processes and integrates them. It is not necessarily data that are entered in these cells but rather brief analytic jottings.

For example, in Barry's fourth data pool at ages 18.5–23.3, one of the entries in the INCREASE/EMERGE cell was: "searching for 'artful living'." Previous data pools in the INCREASE/EMERGE cells included references to his "additional theatre-viewing experiences" from "parental involvement in nurturing his theatre interest," and then "performance opportunities" under "mentorship from theatre teachers," which led to an "attitude 'renaissance'" and a "passion" for the art form. But at ages 18.5–23.3, theatrical performance was no longer part of Barry's life course and thus no reference to it is included in the INCREASE/EMERGE cell for this particular matrix page. (It is, however, noted in the DECREASE/CEASE and MISSING cells.)

Of course, it would be foolish to isolate and strictly compare INCREASE/EMERGE data only with other INCREASE/EMERGE cells. What increases or emerges in a participant's life course must also be compared and considered with other change processes—what has decreased, what is missing, what epiphanies occurred, and so on. If you haven't already, notice that there are dashed and not solid lines separating each cell in the matrix. This was a deliberate artistic and methodological choice to suggest that "the *ocean*—not the landscape—of longitudinal qualitative data knows no boundaries and has open access to flow where needed" (Saldaña, 2005, p. 7). Thus, one entry in this second row's INCREASE/EMERGE cell includes the analytic jotting:

- Barry's new search for "artful living": an attempt to fill an emotional void left by nonparticipation in theatre.

This inferential linking is just one example of how the researcher compares and synthesizes qualitative data across cells in the same row.

Now that the descriptive cells have been discussed, we examine how these observations of difference and thus change are nested within a broader social scheme.

CONTEXTUAL/INTERVENING CONDITIONS INFLUENCING/AFFECTING CHANGES ABOVE

This portion of the matrix is reflective space for jotting how, how much, in what ways, and why the descriptive observations in the top portion of the matrix may have occurred (Saldaña, 2003, p. 161). This is where analysis begins to transform and transcend the descriptive field observations into deeper insight as to individual participant agency, sociocultural factors and environments, historic events leading to the time period, and other major contributors that influence and affect change.

The difference between a contextual and an intervening condition is a matter of researcher interpretation, but let me explain briefly how I distinguish between the two. Contextual conditions are the "givens" of everyday social life—our established environments, structures, processes, and actions that contain and compose the repertory of human routines. Contextual conditions also include our personal "givens"—gender, ethnicity, habits, etc. A child learning addition, a church's congregation worshipping on Sunday morning, an Hispanic social service agency servicing its community, and Barry attending high school, are examples of contextual conditions. Yet even within these everyday contexts, change can happen: the child learns to add one digit figures then progresses onto subtraction; the congregation welcomes new members into their church body and increases in size; the social service agency increases its voluntary staff by two when Hispanic interns join the team for a year; and Barry takes typical school courses and works toward earning a high school diploma.

But when a condition is "perceived as a purposeful, unanticipated, or significant action, structure, or process that influences and affects participant change through time" (Saldaña, 2003, p. 162), the contextual has now become the intervening. "Since contexts are contextual, some conditions initiate greater change than others" (ibid.). When a child learns to add quicker than peers through an experimental teaching approach to mathematics using computers; when the congregation increases in size because a neighboring church's pastor has become involved in a sex scandal; when the Hispanic social service agency loses one of its primary sources of financial support due to their perception of racism in local government and must now rely more on voluntary staff assistance; and when Barry enrolls in theatre courses and participates in extracurricular play productions under the direction of a nurturing mentor, which eventually motivate him to cease illegal drug use; these are examples of intervening conditions and their consequent influences and affects on participants.

"Influences and affects" is my qualitative replacement for the positivist construction of "cause and effect." First, "influences" suggests that there are multiple, networked interactions and complex interplay between and among conditions that drive change. Second, the consequences of those changes create "affects"—multiple outcomes that cognitively, emotionally, and sometimes physically affect, in rippling fashion, participants and their social environments.

This horizontal row is composed of cells that ask you to inductively and/or deductively infer from the descriptive data above the conditions that led to the observations to date. Like the row above this one (DIFFERENCES ABOVE FROM PREVIOUS DATA SUMMARIES), the task is for the researcher to compare and synthesize qualitative data across cells.

INTERRELATIONSHIPS

This cell includes researcher reflections that answer the question, *Which changes interrelate through time?* Interrelationships ("correlations" in quantitative parlance) are researcher-constructed perceptions of how individual phenomena, actions, social structures, change processes and so on, share various types of jointure. Examples may include: the complex interaction and interplay of social institutions (family, community, school, peers, media, etc.) influencing and affecting a young person's attitude, value, and belief systems; a teacher's strategic use of voice (volume, rate, pitch, etc.) in varying qualities to influence and affect student dynamics and thus create consciously-desired learning environments; and Barry's immersion in illegal drug use, the church, and theatre in varying degrees and overlapping phases from ages 12 through 26 as quests for spiritual fulfillment.

Each cell thus far contains essentialized data. And in my originally designed matrix I included the parenthetical direction in the INTERRELATIONSHIPS cell: "circle and connect above, then analyze," meaning: draw lines to link the cells or entries that correlate, comparable to a connect-the-dots drawing. Later I learned that

> this proved to be both ridiculously bravado advice and a nearly impossible task. Interrelationships are human constructions of immense complexity … and my own mind saw the intricate webbing of interaction and interplay between virtually *all items* in the matrix. As Barry's mother herself told me in an interview when I asked her to identify where Barry's interest in theatre came from, the influences were "a great big conglomeration" of factors and experiences that were "just very tightly interwoven. I don't even know how we can separate them." … As I scanned the matrix, I didn't even know how I could *connect* them. (Saldaña, 2005, pp. 14–15)

I've observed that, whether it be the way data are strategically logged in the matrix or some ironic master plan of life comparable to a grand unification theory, there is interrelationship between increases and decreases; between the constant and the missing; between and among the cumulative and contextual; and between and among any other possible and infinite cell combinations. But how do you know where to begin and when to stop? Not everything that interrelates from your perspective may be valid, and any assertions of causality stand on shaky ground in observational studies. Interview data from participants on the influences and affects on their lives will supplement and better inform researcher speculation about connecting chains of action and environmental factors at work:

> The conceptual process for analyzing both interrelationships and change in longitudinal qualitative data is like observing various colored dyes mixing together in the natural currents of water. . . . The eventual discovery of interrelationships between ponds and pools of data is one that emerges after multiple readings of the corpus, data reduction into visual displays, and extended reflection on all possible connections and overlap. (Saldaña, 2003, p. 168)

CHANGES THAT OPPOSE/HARMONIZE WITH HUMAN DEVELOPMENT/SOCIAL PROCESSES

This cell includes researcher reflections that answer the question, *Which changes through time oppose or harmonize with natural human development or constructed social processes?* The observations you make of social life may or may not harmonize with or conform to existing research in your discipline. Examples may include: a first-year, inner-city teacher's classroom management difficulties, comparable to other teachers' experiences in similar settings during the early stages of their professional development; a not-for-profit company's success, whose unexpected patterns of growth and development contradict previous observations and research in established organizational studies; and Barry's epiphanic calling to serve God through the ministry, an experience similar to other spiritual leaders' awareness of their life purpose.

Your literature review and previous research experiences about the current study will inform you whether your longitudinal observations align themselves with established social processes, or whether they oppose what may be considered normative. Unique cases may present qualitative trajectories that contradict or supplement long-standing findings from developmental or organizational studies. Both contextual and intervening conditions, coupled with hypothesized interrelationships (and, in some cases, participant or conceptual rhythms, discussed below), weave together to create a complex tapestry of influences and affects.

This portion of the matrix asks you to relate your own study's observations with the classic and current research in your field to assess how participants stay or stray from the anticipated courses of development and action. Such reflections can support, build on, refute, or revise the extant literature. Analyzing an individual case study or small group of participants in a single site, albeit long-term, may not generate enough data for a convincing argument of transferability to other settings. A larger number of participants and multiple observation sites more convincingly support any assertions of generalizability and theory construction for broader social contexts.

PARTICIPANT/CONCEPT RHYTHMS

The PARTICIPANT/CONCEPT RHYTHMS cell includes researcher reflections that answer the question, *What are participant or conceptual rhythms (phases, stages, cycles, etc.) through time?* Depending on the nature of your research question or study's goals, you may observe that social life and its accompanying changes can be "apportioned into theoretical periodicities of human action" (Saldaña, 2003, p. 169). These periodicities can be constructed as phases, stages, cycles, or other forms and combinations of time-based clusters. Barry himself labeled some of the phases and stages of his own life course as "a dead period," "romantic," the search for "artful living," and an intense period of "academic rigor."

The intuitive or systematic division of your data corpus into varying time periods for multiple matrices may prompt you to investigate whether each matrix page is a constituent unit of coterminous yet varying participant rhythms, much like time signatures, tempos, and measures are parts of written music. It may not always be possible to state with precision when or under what specific contexts the beginning and ending of a phase, stage, or cycle are initiated and concluded. Nevertheless, you should develop a label that encases and describes the rhythmic cluster, with theoretical explanations for how, why, or in what ways a participant enters the apportioned period and transitions from one through the next.

PRELIMINARY ASSERTIONS AS DATA ANALYSIS PROGRESSES

The largest matrix cell is critical space for the longitudinal researcher to "think out loud." This cell includes reflections on how everything above it blends together and (literally) trickles down. However, memo-generation and assertion development (Erickson, 1986) are not reserved exclusively for the final stage of the analytic cycle. Whenever a connection, insight, or even a question occurs to you, no matter at what phase or stage of data entry or analysis, write it down in this cell. Examples from the study with Barry include:

- Barry has always had a rigorous physical outlet in one form or another: football, acting, weightlifting, rock climbing.
- Throughout adolescence, Barry has exhibited high inter- and intrapersonal intelligences, traits Howard Gardner says are conducive to performers and the clergy.
- Ironically, drugs are still playing a role in Barry's life—before, they were illegal and recreational for his depression; now, for his bipolarity, the drugs are prescription—to help him with his depression.

You may find yourself scanning the matrix cells vertically, horizontally, diagonally, systematically, and even randomly to synthesize the data and find answers as if this were some type of elusive "seek and find" word game. My guidance is, "Whatever works." I have concluded that the "longitudinal qualitative researcher's analytic process is neither completely linear nor holistic. It is iteratively—if not erratically—cumulative and serendipitous in knowledge building" (Saldaña, 2005, p. 15).

I prefer to work with hard copy and pencil when I fill in these matrix pages, rather than enter the data into a word-processed file. This means I have to erase and rewrite, rather than delete and rewrite, my assertions in progress. And assertions are almost always in progress, for as data are gathered and entered into future matrix pages, these observations may put previously developed assertions in new perspectives and contexts. Barry's psychological counseling at age 12 and his two suicide attempts during adolescence did not overtly foreshadow his formal diagnosis of bipolarity at 18 years of age. But once this diagnosis was learned, his tragic past and actions "made sense," so to say, and the earlier-developed assertions about his mental and emotional conditions now required revision due to newly acquired disconfirming evidence.

THROUGH-LINE

The THROUGH-LINE cell includes researcher reflections that answer the question, *What is the through-line of the study?* The through-line is a word, sentence, paragraph, and/or extended narrative that captures the essence and essentials of a participant's journey and change (if any) through time. The through-line "includes references to time, processual terms, and markers (beginnings, middles, and/or endings of the journey at various locations through time)" (Saldaña, 2003, p. 170). Examples include:

- From his sophomore through senior years in high school, Barry gradually interchanged the insufficient spiritual fulfillment he received at church with the more personal and purposeful spiritual fulfillment he experienced through theatre (Saldaña, 2003, p. 154).
- *He ascends*. From ages five through twenty-six, Barry has sought *ascension* in both literal and symbolic ways to compensate for and transcend the depths he has experienced throughout his life course (Saldaña, 2005, p. 18).

A through-line tends to evolve toward the end of the analytic cycle, but it is not necessarily the final required outcome from qualitative data analysis. The through-line is a summary statement—a section heading, topic sentence, or theme, if you will—that centralizes the narrative of a longitudinal qualitative study. This can emerge from researcher reflection about the data, or from the participant's own perceptions (and in his or her own language as an "in vivo" code [Strauss and Corbin, 1998]) about the period under investigation.

Recent scholarship (Clarke, 2005; Kincheloe, 2005) advocates that the "messiness, complexity, and interconnectedness of social life cannot be captured through such reductionist methods and are thus futile endeavors. But … the through-line doesn't negate the complexity of a life course, … [it] distills the ocean of longitudinal qualitative data" (Saldaña, 2005, pp. 18–19). The through-line helps navigate the researcher's journey as he or she writes the final epic of his participants' changes (if any) through time.

2 Final comments

James A. Holstein and Jaber F. Gubrium (2000) in *Constructing the Life Course*, remind us that

> The life course and its constituent parts or stages are not the objective features of experience that they are conventionally taken to be. Instead, the constructionist approach helps us view the life course as a social

> form that is constructed and used to make sense of experience.... The life course doesn't simply unfold before and around us; rather, we actively organize the flow, pattern, and direction of experience in developmental terms as we navigate the social terrain of our everyday lives (p. 182).

The Longitudinal Qualitative Data Summary matrix is offered as an organizational method to chart the flows, patterns, and directions of participant experiences and differences and thus change through time. It is a signature technique that I hope has utility and transferability to your own analytic practice. By no means is this matrix proposed as the perfect or only model available. I encourage you to adapt the contents and format to suit your own particular research project and goals. If qualitative data are fluid, then our instrumentation and displays must exhibit comparable fluidity.

As a theatre artist, I approach most of my research projects dramaturgically—meaning, my observations and analyses of social settings are filtered through a "life as drama" framework. When I analyze participant data from interview transcripts, I can't help but apply my training in character analysis gleaned from acting classes. When I analyze physical environments and participant dress and artifacts, my education in the principles of theatrical design emerges. And when I observe social life performed in my presence, I cannot help but apply my directing and playwriting experiences into the collection and analysis of ethnographic data.

Each site visit is a scene or act of a (very long) play. Participants are like characters, each one with objectives, obstacles, tactics, flaws, and emotions. Each one develops and interacts with other characters who have anywhere from inconsequential to significant impact on their lives. But their futures are not scripted and predetermined by a playwright; their lives are improvisationally lived. Yes, there is routine and constancy and hopefully stability to their lives. But nested within and breaking through these repetitive patterns are spontaneous actions that force them—and us—to deal with life as it is complexly and challengingly lived. Sometimes we can predict with statistical precision, based on selected variables of interest, what pathways or outcomes lie ahead. But most times the future is unknown. "What happens next?" is the central question in the audience's minds as we watch an engaging play, film, or television story unfold before us. "What happens next?" is also the driving question for longitudinal observational fieldwork and its concurrent qualitative data analysis.

Glossary

Assertions Descriptive, analytic, or interpretive summary statements, derived from qualitative data, of phenomena, participant actions, or social environments.

Change processes Category-based transformations through time (e.g.,"increase," "decrease,""idiosyncratic").

Dynamics The dimensions and variability of qualitative data, expressed through carefully-selected language.

Matrix An interrelated, multicelled chart for qualitative data entry and analysis.

Qualitative observational data Written fieldnotes gathered primarily from participant observation in naturalistic social settings.

References

Clarke, A. E. (2005). *Situational Analysis: Grounded Theory After the Postmodern Turn*. Thousand Oaks: Sage.

DeWalt, K. M., and DeWalt, B. R. (2002). *Participant Observation: A Guide for Fieldworkers*. Walnut Creek: AltaMira Press.

Emerson, R. M., Fretz, R. I. and Shaw, L L. (1995). *Writing Ethnographic Fieldnotes*. Chicago: University of Chicago Press.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (ed.), *Handbook* *of Research on Teaching*, 3rd edn, pp. 119–161. New York: Macmillan.

Fullan, M. (2001). *The New Meaning of Educational Change*, 3rd edn. New York: Teachers College Press.

Holstein, J. A. and Gubrium, J. F. (2000). *Constructing the Life Course*, 2nd edn. Dix Hills, NY: General Hall.

Huber, G. P. and Van de Ven, A. H. (eds) (1995). *Longitudinal Field Research Methods*. Thousand Oaks, CA: Sage.

Kelly, J. R. and McGrath, J. E. (1988). *On Time and Method*. Newbury Park: Sage.

Kincheloe, J. L. (2005). On the next level: Conceptualization of the bricolage. *Qualitative Inquiry*, 11(3): 323–350.

Levine, R. (1997). *A Geography of Time*. New York: Basic Books.

Lofland, J., Snow, D., Anderson, L. and Lofland, L. H. (2006). *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis*. Belmont: Thomson Learning.

Menard, S. (1991). *Longitudinal Research*. Newbury Park: Sage.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative Data Analysis*, 2nd edn. Thousand Oaks, CA: Sage.

Royce, A. P. and Kemper, R. V. (2002). Long-term field research: Metaphors, paradigms, and themes. In R. V. Kemper and A. P. Royce (eds), *Chronicling Cultures: Long-term Field Research in Anthropology*, pp. xii–xxxviii. Walnut Creek: AltaMira Press.

Saldaña, J. (1995). "Is theatre necessary?" Final exit interviews with sixth grade participants from the ASU longitudinal study. *Youth Theatre Journal*, 9: 14–30.

Saldaña, J. (1996). "Significant differences" in child audience response: Assertions from the ASU longitudinal study. *Youth Theatre Journal*, 10: 67–83.

Saldaña, J. (1997). "Survival": A white teacher's conception of drama with inner-city Hispanic youth. *Youth Theatre Journal*, 11: 25–46.

Saldaña, J. (1998a). Ethical issues in an ethnographic performance text: The "dramatic impact" of "juicy stuff". *Research in Drama Education*, 3(2): 181–196.

Saldaña, J. (1998b). "Maybe someday, if I'm famous": An ethnographic performance text. In J. Saxton and C. Miller (eds), *Drama and Theatre in Education: The Research of Practice, the Practice of Research*, pp. 89–109. Brisbane: IDEA Publications.

Saldaña, J. (2003). *Longitudinal Qualitative Research: Analyzing Change through Time*. Walnut Creek: AltaMira Press.

Saldaña, J. (2005). Coding qualitative data to analyze change.Unpublished manuscript.

Spradley, J. P. (1980). *Participant Observation*. Fort Worth: Harcourt Brace Jovanovich.

Strauss, A. and Corbin, J. (1998). *Basics of Qualitative Research*, 2nd edn. Thousand Oaks: Sage.

Sztompka, P. (1993). *The Sociology of Social Change*. Oxford UK: Blackwell.

Tudge, J. and Hogan, D. (2005). An ecological approach to observations of children's everyday lives. In S. Greene and D. Hogan (eds), *Researching Children's Experience: Approaches and Methods*, pp. 102–122. Thousand Oaks: Sage.

This page intentionally left blank

Chapter 20

Configural frequency analysis of longitudinal data

Alexander von Eye and Eun Young Mun

1 Introduction

In this chapter, first, the perspectives are discussed that researchers take when applying configural frequency analysis (CFA). Specifically, when applying CFA, the focus is on individuals who differ in profiles or patterns instead of relationships among variables. A tutorial of CFA is given. In the section on longitudinal methods of CFA, four ways of data analysis are presented. The first involves the analysis of group-specific temporal patterns. Using this method, one can identify the patterns of development that groups differ in. This method is compared with odds ratio analysis. The second longitudinal CFA method allows one to analyze trends over time. Specifically, the shape of change curves are analyzed, e.g., the linear and the quadratic trends. It is discussed that the simultaneous analysis of trends that differ in order (e.g., linear and quadratic) can lead to cells in cross-classifications that contain structural zeros. The third CFA method for the analysis of longitudinal data involves taking a priori probabilities into account. The fourth method allows one to compare specific transition patterns with each other. Empirical examples are given using data from a study on the development of aggression in adolescence.

Models for longitudinal data contain parameters that are unique in the sense that they are not part of models for cross-sectional data. These parameters concern, for instance, the linear trend, the acceleration of the linear trend, the change in acceleration, transition patterns, patterns of constancy, and information concerning the association structure of variables and its change over time. *Configural frequency analysis* (CFA; Lienert and Krauth, 1975; von Eye, 2002; von Eye and Gutiérrez-Peña, 2004) is a method for the examination of multivariate categorical data. CFA allows one to study any parameter of longitudinal data, in particular if it can be placed in the context of categorical data.

This chapter presents a selection of methods of longitudinal CFA. Specifically, we describe and illustrate (1) CFA of differences; (2) CFA of trends; and (3) CFA of symmetry in change. Before covering these facets of CFA of longitudinal data, we present an introduction to the concepts and methods CFA.

2 CFA—a tutorial

Consider the cross-classification of d categorical variables. The cells of this cross-classification contain the observed frequencies of each pattern of variable categories, called *configuration*. Using log-linear models, cross-classifications can be analyzed, focusing on the joint frequency distribution, the dependency structure, and the association structure of the d variables (Goodman, 1984). CFA offers an alternative perspective. It allows one to examine individual cells of a cross-classification. Let the observed cell frequency of a cell be denoted by m_r and the corresponding estimated expected frequency by e_r, where r goes over all cells of the cross-classification. Then, based on the comparison of m_r with e_r, CFA states either that

- Cell r constitutes a *CFA type* if $m_r > e_r$, or
- Cell r constitutes a *CFA antitype* if $m_r < e_r$.

If $m_r = e_r$, Cell r constitutes neither a type nor an antitype. The null hypothesis that must be rejected for these decisions is

$$H_0\colon E[m_r] = e_r,$$

where $E[\ldots]$ indicates the expectancy, and m_r and e_r are defined as above. The expected frequency e_r is estimated under a model called base model. This model is specified such that it allows one to interpret types and antitypes in a clear-cut way (more detail follows later). In many cases, the base model is a log-linear model. In particular in the context of longitudinal data analysis other base models have been discussed also. This chapter will discuss two examples of such models.

2.1 Testing cell-wise hypotheses in CFA

Let the d variables, $X_1, \ldots, X_d$ be crossed to form a contingency table with $R = \prod_{i=1}^{d} c_i$ cells, where c_i is the number of categories of the ith variable. Let the probability of Cell r be π_r, with $r = 1, \ldots, R$. The frequency with which cell r was observed, is m_r. The probabilities of the R cell frequencies depend on the sampling scheme of the data collection (von Eye and Schuster, 1998; von Eye, Schuster and Gutiérrez-Peña, 2000). The typical sampling schemes are *multinomial* and *product-multinomial*. Mixed schemes are applied also. If sampling is multinomial, cases are randomly assigned to cells of a table, and there are no constraints except that the sample size is given. If sampling is product-multinomial, marginal frequencies are fixed, and assignment has to proceed such that the marginal frequencies are reproduced.

In most cases, sampling is multinomial and we obtain

$$P(m_1, \ldots, m_R \,|\, N, \pi_1, \ldots, \pi_R) = \frac{N!}{m_1! \ldots m_R!} \prod_{r=1}^{R} \pi_r^{m_r}$$

with $\sum \pi_r = 1$ and $\sum m_r = N$. Because of the multinomial sampling scheme, m_r is binomially distributed with

$$P(m_r \,|\, N, \pi_r) = \frac{N!}{m_r!(N - m_r)!} \pi_r^{m_r} (1 - \pi_r)^{N - m_r}$$

To test hypotheses about individual cells, one can use the binomial distribution, and one obtains

$$B_{N,p}(x) = \sum_{j=0}^{x} \frac{N!}{j!\,(N-j)!} p^j \; (1-p)^{N-j}$$

with $0 \leq \times \leq N$. A number of alternative tests has been proposed. The most popular among these are the X^2 component test, the z test, and, for product-multinomial sampling, exact and approximate hypergeometric tests (Lehmacher, 1981). The X^2 component test is where, as before, r goes over all cells of the cross-classification. This test comes with $df = 1$, and it is well known that $\sqrt{X^2} = z$, the standardized residual.

Lehmacher's asymptotic hypergeometric test can be described as follows. Considering that the denominator of X^2 tends to underestimate the variance of the normal distribution when the model describes the data well (in this case, $\sigma^2 < 1$), Lehmacher (1981) derived the exact variance for Cell r

$$\sigma_r^2 = Np_r\,[(1 - p_r - (N-1)(p_r - \tilde{p}_r)]$$

where N is the sample size, p_r is the probability of being assigned to Cell r, and $\tilde{p}_r$ is estimated to be

$$\tilde{P}_r = \frac{1}{(N-1)^d} \prod_{j=1}^{d} (N_{jk_j} - 1)$$

where k_j indexes the univariate marginal frequencies of variable J. The probability p_r is estimated using the CFA base model. In standard CFA, under product-multinomial sampling, the base model is the model of variable independence, i.e., the main effect model. Lehmacher's test statistic is

$$Z_{L,r} = \frac{m_r - e_r}{\sigma_r}$$

This test statistic is asymptotically standard normally distributed. Because $p_r > \tilde{p}_r$, Lehmacher's z will always be larger than $\sqrt{X^2}$, and the test has more power to classify configurations as constituting types or antitypes than $\sqrt{X^2}$. To prevent non-conservative decisions, Küchenhoff (1986) recommended using a continuity correction.

A number of theoretical and simulation studies has been undertaken to identify the best performing among the many tests that have been proposed for CFA (e.g., Indurkhya and von Eye, 2000; von Eye and Mun, 2003; von Weber, Lautsch and von Eye, 2003a; von Weber, von Eye and Lautsch, 2004). The results of these comparisons are that

$$X^2 = \frac{(m_r - e_r)^2}{e_r}$$

1. Whenever possible, exact tests are to be preferred. These are, for instance, the binomial and the exact hypergeometric tests.
2. Of the asymptotic tests, none performs always the best. However, the X^2 component (and the z-) test, and the two procedures proposed by Perli, Hommel, and Lehmacher (1985) and by von Weber et al. (2004) cover 90% of the best solutions. This applies to both keeping the α and reasonably low β-levels.
3. When samples are small, antitypes are hard or impossible to detect. When samples are large, antitypes are more likely to be detected. This applies to all tests with the exception of Lehmacher's test for very small tables, for which it detects equal numbers of types and antitypes.

2.2 Protecting α

The typical application of CFA is in an exploratory context. In this context, it is unknown whether types and antitypes will emerge, and where they may emerge. Therefore, all R cells in a table are examined. Clearly, the number of tests can be large when the size of a table is large. The factual significance level α can be guaranteed to be equal to the nominal level only for the first of a series of tests on the same sample.

There are two reasons why, for a large number of tests, the factual and the nominal levels α differ from each other. The first reason is the *mutual dependence of multiple tests*. The extreme case is a 2 × 2 table which is analyzed under the main effect model. von Weber, Lautsch and von Eye (2003b) showed that, under these conditions, the results of all CFA tests after the first test are completely dependent upon the result of the first. When the size of the tables increases (both in number of dimensions and number of variable categories), this dependency becomes less pronounced. However, tests never become completely independent upon each other (see also Steiger, Shapiro and Browne, 1985). Because of this interdependency, the factual level α can represent a severe underestimation of the nominal level α.

The second reason, exacerbating the first, is that of *capitalizing on chance*. In the present context, this means that because researchers expose themselves repeatedly to the risk (of size α) of committing an α-error, the probability increases beyond α with each test. von Eye (2002) presents the following example. If a researcher examines each of the 27 cells in a 3 × 3 × 3 table under the nominal level α, the probability of committing a Type I error three times is $p = 0.1505$, even if the tests are, in truth, independent. In other words, with probability 0.1505, three of the 27 type/antitype decisions will be wrong.

Because of the great danger of making wrong decisions in CFA, protection of α has become routine in CFA applications. For exploratory applications, Perli, Hommel, and Lehmacher (1985) proposed protecting the *local level* α. This method guarantees that, for each cell-wise test, the factual α does not exceed the nominal level α. However, protection of α comes with a price: the decisions concerning the cell-wise null hypotheses become far more conservative. In fact, unless samples are really large, types and antitypes may become extremely hard to detect.

A number of procedures to protect α has been proposed. The most popular procedure, known as *Bonferroni adjustment*, is also the most conservative. Let the number of tests be R. Then, the Bonferroni procedure requires that two conditions be met. First, the sum of all individual α_r does not exceed the nominal α, i.e., $\sum_{r=1}^{R} \alpha_r \leq \alpha$, where r goes over all tests. Second, the procedure requires that all α_r be equal, or $\alpha_r = \alpha^*$, for all r, where α^* is the adjusted significance threshold. The adjusted threshold that meets both requirements is $\alpha^* = \alpha/R$.

To illustrate the effect of this adjustment, consider the example with the 3 × 3 × 3 table again. For 27 tests, the adjusted significance threshold is no longer $\alpha = 0.05$. Instead, it is $\alpha^* = 0.05/27 = 0.00185$. The z-score for the nominal α is ±1.96. The z-score for the adjusted α^* is ±2.901.

Because Bonferroni's procedure is so extreme, a number of less conservative procedures has been proposed. Here, we review the procedures proposed by Holm (1979) and by Hommel (1988; 1989). Holm proposed to take into account the number of tests performed before the ith test. Taking this number into account yields a protected α^* that becomes less conservative as the number of tests increases. Specifically, Holm's protected significance threshold is

$$\alpha_r^* = \frac{\alpha}{R - r + 1}$$

where r is the number of the test currently performed, for $r = 1, \ldots, R$. Because the number of tests performed before the rth is taken into account, the order of tests is no longer arbitrary. Therefore, the test statistics must be rearranged in descending order. Thus, the largest test statistic is considered first, the second largest after that, etc. When the probabilities are ordered that can be calculated for the exact tests, the order is ascending. The ordering requires additional work. However, this amount of work is compensated in part by the fact that the testing of configurations concludes as soon as the first null hypothesis can be retained.

To illustrate the less conservative nature of Holm's procedure, we compare it with Bonferroni's procedure. For the first test ($r = 1$), one obtains $\alpha_1^* = \frac{\alpha}{R-1+1} = \frac{\alpha}{R}$. That is, for the first test, the Holm and the Bonferroni procedures require the same threshold. Beginning with the second test, however, the Holm procedure is less restrictive. For $i = 2$, one obtains $\alpha_2^* = \frac{\alpha}{R-2+1} = \frac{\alpha}{R-1}$, which is larger than α^*. Holm's thresholds become increasingly less restrictive until the Rth test, for which one obtains $\alpha_R^* = \frac{\alpha}{R-R+1} = \alpha$.

For two- and three-dimensional tables, Holm's procedure can be made even less restrictive. Here, we present the results for three-way tables. The adjusted α-level is, for $r = 1, \alpha_1^* = \alpha/R$. As for Holm's procedure, this value is the same as for Bonferroni's procedure. However, for $r = 2, 3, 4,$ and 5, one obtains $\alpha_2^* = \ldots = \alpha_5^* = \alpha/(R-4)$. These values are less restrictive than the adjusted levels under the Bonferroni and the Holm procedures. The remaining tests are performed under the same adjusted thresholds as under the Holm procedure.

2.3 Base models for CFA

In the context of modeling, researchers attempt to devise the best-fitting model. Large model-data discrepancies are used as indicators of poor fit, and as hints at ways for model improvement. Using CFA, researchers pursue different goals (Lehmacher, 2000; von Eye, 2002). Specifically, using CFA, researchers attempt to find types and antitypes that are interpretable with reference to a particular base model. The base model must be tailored such that a specific interpretation is possible. Because of this goal, (1) the specification of a base model is neither trivial nor arbitrary, and (2) model-data discrepancies do not lead automatically to a modification of the model. Instead, they lead to interpretable types and antitypes. The two approaches to data analysis, modeling and search for types share the characteristic that model-data discrepancies are model-specific. Therefore, if the same configuration emerges as a type or antitype for more than one base model, it can have quite different interpretations.

von Eye (2004) has proposed a taxonomy of CFA base models (see von Eye and Schuster, 1998; von Eye et al., 2000). This taxonomy classifies CFA base models into four groups: (1) log-linear base models; (2) models with configural probabilities based on known population parameters; (3) models with configural probabilities based on a priori considerations (these models are of particular importance in longitudinal research); and (4) base models that reflect assumptions concerning the joint distribution of the variables under study (von Eye and Bogat, 2004; von Eye and Gardiner, 2004). Mixed base models have been discussed also. In the following sections, we review the four groups of base models.

In general, a CFA base model includes all those variable relationships that are *not of interest* to the researcher. Types and antitypes can then emerge only if the relationships of interest exist. There are three criteria for base models (von Eye and Schuster, 1998).

1. *The base model must allow for unique interpretation of types and antitypes*. Types and antitypes reflect discrepancies between model and data. A model qualifies as a CFA base model if there is only one reason for these discrepancies. Examples of such reasons include interactions, main effects, and predictor-criterion relationships.
2. *The base model must be parsimonious*. That is, the base model must contain as few terms as possible, and these terms must be of the lowest possible order (for methods to reduce the complexity of base models, see Schuster and von Eye, 2000).
3. *The base model must consider the sampling scheme*. When sampling is multinomial, practically every conceivable base model is possible. However, when sampling is product-multinomial, the fixed marginals must be reproduced. This applies in particular when the sampling scheme is product-multinomial in more than one variable. In this case, the base model must reproduce the marginal probabilities in the subtables that result from crossing variables with fixed margins. Base models that do not consider these subtables are, therefore, inadmissible.

Examples follow in the next section.

Log-linear base models for CFA

Log-linear models (Agresti, 2002; Goodman, 1984) allow one to describe the joint distribution, the dependency structure, and the association structure in a cross-classification of categorical variables. Log-linear models are the most popular base models of CFA. There are two groups of log-linear base models. The first is the group of *global base models*. In these models, all variables have the same status. There is no distinction between dependent and independent variables, predictors and criteria, or even separate groups of variables.

There is a hierarchy of global models. Models are higher in the hierarchy as the complexity of variable relationships increases, which the base model takes into account. In ascending order, the simplest base model is the null model. It requires multinomial sampling. This model assumes that no effects exist. The log-linear formulation is this $\log m = \lambda$, i.e., contains only the constant term. Types and antitypes reflect the existence of any effect. However, the nature of these effects is, without further analysis, not determined.

The next-higher model is that of variable main effects. This model is the independence model, also called *model of first-order CFA*. As the name indicates, this model takes main effects into account. Types and antitypes thus can emerge only if the variables interact.

One level higher in the hierarchy is the model of first-order interaction, also called *model of second-order CFA*. This is a hierarchical model in the sense that it also takes into account the main effects of all variables. Types and antitypes can emerge only if second- or higher-order interactions exist. Higher-order models can be considered. However, to the best of our knowledge, there have been no applications of higher-order models.

The second group of log-linear CFA base models contains *regional base models*. In contrast to global models, regional models are used to analyze groups of variables. These groups can be predictors and criteria, or simply two or more groups of variables that are related to each other. In longitudinal research, the groups of variables may be observed at different points in time, and researchers ask whether patterns of categories at the first point in time allow one to predict patterns of categories at the next point in time.

Base models that use known population probabilities

The typical application of log-linear models uses sample information to estimate cell probabilities. However, occasionally, population probabilities are known as, for instance, the incidence rates of cancer, the gender distribution in a particular age bracket, the rate of car accidents in Michigan, or the number of children born to women in Italy. Whenever population probabilities are known, one can ask whether a sample can be assumed drawn from this population. If there are significant discrepancies, one may ask where they are located, and this is the domain of CFA. Types indicate where more cases can be found than expected based on population parameters, and antitypes indicated where the number of cases is smaller than expected based on population parameters. In longitudinal research, population information is rarely available. Therefore, base models that reflect such information have not been used frequently.

Base models based on a priori probabilities

Of particular importance in longitudinal research is the possibility that the rate of change patterns does not only differ empirically but also theoretically, a priori. One frequently-used method of longitudinal CFA targets intraindividual changes in a series of measures by creating variables that indicate, the linear, quadratic, and higher-order (polynomial) trends. When data are continuous, linear trends can be examined using the methods of first differences, i.e., differences between time-adjacent scores. Second differences, i.e., differences between first differences, are used to analyze quadratic trends (details follow below). Of these differences, often just the signs are analyzed to indicate whether there is an increase or decrease, or an acceleration or a deceleration. It has been shown (von Eye, 2002, Ch. 8) that patterns of signs of these differences come with different a priori probabilities. CFA base models can take these into account.

Base models that take distributional assumptions into account

Recently (von Eye and Bogat, 2004; von Eye and Gardiner, 2004), a variant of CFA was proposed that allows one to examine sectors in the multivariate space for which the expected probability was estimated under the assumption of multinormality (Genz, 1992). Types (or antitypes) suggest that a sector contains more (or fewer) cases than expected based on the assumption of multinormality. Thus, CFA can be used as a method for testing the multinormality assumption. The interesting aspect of this approach to testing for multinormality is that results indicate which sectors exactly show significant departures from multinormality. If it can be assumed that samples do come from multinormal populations, one can ask why these sectors show discrepant probabilities, and whether *selective resampling* can be considered to make a sample more representative of the population it was drawn from.

In the following sections, we present methods of CFA for the analysis of longitudinal data. We begin with the analysis of two-wave data.

3 CFA of longitudinal data

In this section, we present a selection of the many methods that have been proposed for the configural analysis of longitudinal data. We begin with a method that allows one to determine whether temporal patterns of the same variable, observed over two or more occasions, are specific to one of two or more groups under study. The second approach to be presented allows one to analyze patterns of trends. The third approach focuses on symmetry tables.

3.1 CFA of group-specific temporal patterns

Consider the case in which a categorical variable was observed repeatedly in two or more groups of respondents. We then can ask whether the transition patterns are group-specific. In the simplest case, we have a dichotomous variable that was observed twice. Let the categories of this variable be labeled with 1 and 2. From two categories, four transition patterns result: 11, 12, 21, and 22. Still dealing with the simplest case, we observe these transition patterns in two groups of respondents. Crossing the group variable with the transition patterns yields the eight patterns 111, 112, 121, 122, 211, 212, 221, and 222. Let the grouping variable be the last of the three variables in the crossing patterns. The first of these eight configurations describes those respondents who show Category 1 on both occasions, and these respondents are from Group 1. Accordingly, Configuration 212 describes those respondents who show Category 2 at the first occasion, Category 1 at the second occasion, and are members of the second group. Table 20.1 illustrates this 2 × 2 × 2 arrangement.

For the following considerations, we assume that no population parameters are known and no a priori probabilities exist. In addition, we assume multinomial sampling. Now, the question as to whether transition patterns are group-specific can be answered in more than one way. The routine way would involve calculating odds ratios (see Rudas, 1998). A first odds ratio is

$$\theta_1 = \frac{p_{111}/p_{121}}{p_{112}/p_{122}}$$

Table 20.1 2 × 2 × 2 Cross-classification for comparison of transition patterns in two groups

	Group 1			*Group 2*	
	Time 2			*Time 2*	
Time 1	m_{111}	m_{121}	Time 1	m_{112}	m_{122}
	m_{211}	m_{221}		m_{212}	m_{222}

where p_{ijk} indicates the probabilities of the eight patterns and the grouping variable is the last in the subscripts. This odds ratio tells us whether the transition from 11 to 12 is observed at a different rate in Group 1 than in Group 2. The null hypothesis according to which the logarithm of the odds ratio (the *log odds*) is equal to zero (even odds), i.e., that the transition rates from 11 to 12 are the same in the two groups, can be tested using the standard normal

$$z_{\log\theta} = \frac{\log\theta}{\left(\sum_{1}^{4}\frac{1}{m_{ijk}}\right)^{0.5}}$$

for $i = 1$, and $j,\ k = \{1,2\}$.

Accordingly, we can ask whether the transition from 21 to 22 is the same in the two groups, and calculate

$$\theta_2 = \frac{p_{211}/p_{221}}{p_{212}/p_{222}}$$

The significance test applies accordingly. In a follow-up step, we can compare these two odds ratios (or these two log odds) by calculating $\Omega = \theta_1/\theta_2$. This ratio tells us whether the group differences in transition rates are the same for the transition from Category 11 to 12 as for the transition from Category 21 to 22. The significance test statistic for log Ω is

$$z_{\log\Omega} = \frac{\log\Omega}{\left(\sum_{1}^{8}\frac{1}{m_{ijk}}\right)^{0.5}}$$

for $i,\ j,\ k = \{1,2\}$.

There can be no doubt that these odds ratios and the ratio of the two odds ratios allow us to analyze interesting and important questions. We learn whether transition patterns differ across groups and whether these differences carry over to other transitions. In addition, odds ratio analysis has important characteristics. For example, odds ratios are *marginal-free*. That is, odds ratios are not affected by extreme marginal distributions, as long as the ratios stay unchanged.

However, CFA allows one to ask additional questions. Specifically, CFA allows one to *ask whether the groups differ in individual patterns*. For example, we can ask whether the two groups in the present example differ in the rates in which they show transition pattern 12. To answer this question (and the corresponding ones, for the remaining four transition patterns), we need to develop a base model. This model must be specified such that types and antitypes can emerge only if the groups differ in the rates with which they exhibit a transition pattern. In other words, the correlation between the scores observed at the first and the second occasions must not lead to types and antitypes. In addition, the occurrence rates of the variable categories must not lead to types and antitypes either.

The log-linear base model with these characteristics takes into account the following effects:

- Main effects of all three variables (Time 1, Time 2, and Group); taking the main effects into account has the effect that differences in occurrence rates of variable categories and differences in group sizes do not lead to the emergence of types and antitypes.
- Interaction between Time 1 and Time 2; taking this interaction into account has the effect that autocorrelations will not lead to types and antitypes. This applies accordingly if data from more than two occasions are analyzed.

In contrast, none of the interactions of the grouping variable and the two observation points is taken into account. If one or more of these interactions exist, they will manifest in the form of types and antitypes. In brief, the hierarchical log-linear base model of choice is

$$\log m = \lambda + \lambda_{ij}^{T1,T2} + \lambda_{k}^{G}$$

The first term on the right side of the equation is the constant. The second term is the interaction between the two observation points, T1 and T2. Because the model is hierarchical, the main effects of T1 and T2 are implied. The last term is the main effect of the grouping variable, G. In bracket notation, this model is [T1, T2][G]. In this model, only the interactions between G and the two observation points are missing. These are reflected in the terms [T1, G], [T2, G], and [T1, T2, G]. If these interactions exist, the base model is rejected and types and antitypes exist. It is important to note that the CFA base model does take the main effects into account. It is thus *marginal-dependent*. When data from more than two occasions are analyzed, the base model takes the highest possible interaction among observation points into account. For example, for the four data waves T1, T2, T3, and T4, the base model is [T1, T2, T3, T4].

Clearly, the only effects that, for the above base model, can lead to the emergence of types and antitypes, are those that link T1 and T2 with G. These are the effects [T1, G], [T2, G], and [T1, T2, G]. There is no other way to contradict this base model, and the first, the most important condition for CFA base models is fulfilled. If the base model is rejected, we can examine cells for local model-data discrepancies.

In the current context, we will proceed in a slightly different way than inspecting each individual cell. Instead, we inspect pairs of cells. This procedure allows us to compare the two groups of respondents. We thus perform a 2-group repeated measures CFA – a model not discussed thus far in the CFA literature (see von Eye, 2002).

Elements of 2-group CFA

The function of the above log-linear model is to estimate the expected cell frequencies. These will, under the null hypothesis discussed in the introduction, be compared with the corresponding observed frequencies. In 2-group CFA, this is done after pairs of cells are arranged in the format of a 2 × 2 cross-tabulation. The columns of this table are constituted by the two groups under study. One of the rows contains the two observed frequencies of the transition pattern used to compare the two groups, the other row contains the observed frequencies summed over the remaining transition patterns. Thus, each comparison is performed *in the context of the whole table*. As was indicated above, the expected cell frequencies are estimated using a log-linear base model. Table 20.2 illustrates this arrangement (see von Eye, 2002, p. 175).

Cross-classifications as the one depicted in Table 20.2 can be analyzed using, for instance, the exact Fisher test, odds ratios, the X^2-test, or a number of z-approximations. When the test suggests that the null hypothesis can be rejected, there is a relationship between group membership and transition pattern or, in other words, the transition pattern is group-specific. If a pattern is group-specific, it is said to *constitute a discrimination type*.

Data example

For the following example, we use data from a study on the development of aggression in adolescents (Finkelstein, von Eye and Preece, 1994). In this study, 38 boys and 76 girls in the UK were asked to respond to an aggression questionnaire in 1983, 1985, and 1987.

Table 20.2 2 × 2 table for 2-group CFA

Configuration	*Group*		*Row totals*
	G1	*G2*	
Comparison pattern *ij*	m_{ijG1}	m_{ijG2}	$m_{ij.}$
All others combined	$m_{..G1} - m_{ijG1}$	$m_{..G2} - m_{ijG2}$	$N - m_{ij.}$
Column totals	$m_{..G1}$	$m_{..G2}$	N

The average age at 1983 was 11 years. One of the dimensions of aggression examined in this study was verbal aggression against adults (VAAA). In the present example, we ask whether the development of VAAA from 1983 to 1985 is gender-specific.

To analyze the data, we dichotomized the two measures of verbal aggression, VAAA83 and VAAA85 at the median, with 1 indicating below and 2 indicating above average scores, and crossed the dichotomized variables with gender (1 = males, 2 = females). This resulted in the VAAA83 × VAAA85 × Gender cross-classification. The expected frequencies were estimated using the log-linear base model

$$\log m = \lambda + \lambda_{ij}^{VAAA83,VAAA85} + \lambda_k^G$$

For the individual tests, we used the z-approximation of the binomial test, and α was protected using the Bonferroni procedure. The adjusted α was $\alpha^* = 0.0125$. Table 20.3 displays the results of 2-group CFA.

The results in Table 20.3 show that the first three transition patterns (from 1 to 1, from 1 to 2, and from 2 to 1) contain slightly more male adolescents than one would expect from the 1:2 ratio in the sample. Each of these transition patterns contains at least one indication of below average verbal aggression. However, none of these three differences is strong enough to allow for a significant discrimination between female and male respondents. However, transition pattern 22 does indicate a significant gender difference in transition pattern. It constitutes a discrimination type, indicating that significantly more female than male respondents considered themselves above average in verbal aggression at both ages 11 and 13 years.

This result cannot be reproduced using odds ratio or log-linear analysis. The hierarchical log-linear model [VAAA83, VAAA85][G, VAAA85] explains the data well (LR-$X^2 = 2.75$; $df = 2$; $p = 0.25$). Compared to the above base model, only the second term was added. However, this model only states that the auto-association between the two verbal aggression scores is significant, and that the second verbal aggression measure is associated with Gender. The model does not allow one to talk about transition patterns at the level of individual patterns and their relationship with the classification variable, Gender. Therefore, CFA and log-linear modeling can be seen as complementing each other, while answering different research questions.

3.2 CFA of patterns of trends

Differences between time-adjacent observations tell us whether a later observed score is greater

Table 20.3 2-group analysis of patterns of the development of verbal aggression

Configuration	m_{ijk}	*Statistic*	*p*	*Type?*
111	15			
112	22	1.132	.129	
121	8			
122	14	.336	.369	
211	9			
212	11	1.219	.111	
221	6			
222	29	−2.441	.007	Discrimination type

than an earlier observed score. More specifically, we calculate $\Delta y_i = y_{i+1} - y_i$. If $\Delta y_i > 0$, then the score at Time $i+1$ is greater than the score at Time i. Scores Δy_i are called *first differences*. Positive first differences indicate an increase in scores over time.

Accordingly, *second differences* can be defined. Given the first differences, Δy_i, one can calculate $\Delta^2 y_k = \Delta y_{k+1} - \Delta y_k$. Second differences indicate whether the trend in the first differences changes over time. That is, second differences tell us whether the first differences are accelerated or decelerated over time. Curvature can be examined by calculating third and higher-order differences.

The method of differences has a number of important characteristics:

1. The data analyzed with this method must be at the interval or ratio scale levels.
2. First differences describe linear trend (for the use of differencing methods in time series analysis, see e.g. Chapters 34 and 36 in this volume); in this characteristic, they are comparable to first-order polynomials.
3. Second differences describe changes in the linear trend; in this characteristic, they are comparable to quadratic polynomials; this applies accordingly to higher-order differences and polynomials.
4. If the first differences are constant, the series of scores can be described by a linear regression line.
5. If the first differences vary, but the second differences are constant, the series of scores can be described by a quadratic function; this applies accordingly to higher-order differences.
6. Analyzing first differences reduces the number of available scores in the series by one; analyzing second differences reduces the number of available scores by two, etc.
7. Data points must be equidistant.
8. For k repeated observations, testing change parameters for polynomials of order up to $k-1$ becomes possible.

The method of differences just described is called *method of descending differences* if later observations are subtracted from earlier differences. The *method of ascending differences* involves subtracting earlier observations from later ones. The *method of central differences* subtracts scores from a common reference point, typically the mean or median.

Data example

In the following application example, we use the data from the Finkelstein et al. (1994) study again. In addition to verbal aggression, aggressive impulses were assessed. We now ask whether there are types or antitypes of change patterns in linear and quadratic trends in aggressive impulses. In a first step, we calculate the difference scores. We obtain the variables DAI1 = AI85 − AI83, DAI2 = AI87 − AI85, and DAI2ND = (AI87 − AI85) − (AI85 − AI83), where AI83, AI85, and AI87 are the aggressive impulse scores for the years 1983, 1985, and 1987, respectively.

The reason why we include the quadratic trend in addition to the linear one is that, at least for some of the adolescents, the changes in aggressive impulses cannot be satisfactorily described by only linear trends. Figure 20.1 shows all 114 trajectories. Clearly, some of the trajectories are U-shaped, whereas others are inversely U-shaped.

In the next step, the first and second differences were dichotomized at the zero point, with 2 indicating positive differences and 1 indicating negative differences. The resulting two values for the first differences between 1985 and 1983 and between 1987 and 1985 thus indicate a positive linear trend for a score of 2 and a negative linear trend for a score of 1. Accordingly, the resulting values for the quadratic trend indicate a U-shaped trend for a score of 2 (acceleration) and an inversely U-shaped trend for a score of 1 (deceleration).

Using CFA, we now ask whether particular patterns of first and second differences occur

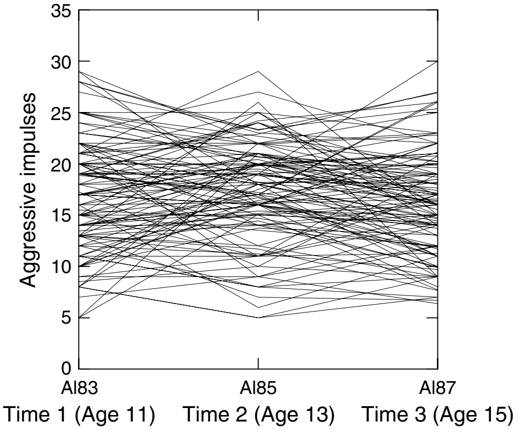


Figure 20.1 Parallel coordinate display for self-perceived aggressive impulses in the years 1983, 1985, and 1987

more often or less often than expected. To specify expectation, we use the base model of variable independence, i.e.,

$$\log m = \lambda + \lambda_i^{DAI1} + \lambda_j^{DAI2} + \lambda_k^{DAI2ND}$$

From this base model, types and antitypes can emerge if any of the pairwise or three-way interactions exist. We adjust α using the Bonferroni procedure which results in $\alpha^* = 0.00625$, and we use the z-test. Table 20.4 displays results of this first-order CFA.

The results in Table 20.4 suggest that 2 types and 2 antitypes exist. The first type is constituted by Configuration 122 (see also Figure 20.2). This pattern is characterized by a decrease in aggressive impulses from 1983 to 1985 that is followed by an increase. This pattern is complemented by a U-shaped quadratic trend; 26 respondents displayed this pattern, but only about 9 had been expected. The second type (Configuration 211) shows the opposite pattern. An increase in aggressive impulses is followed by a decrease. This pattern is complemented by an inversely U-shaped quadratic trend. 36 adolescents were observed to display this pattern – about 1.5 times as many as expected.

The antitypes are constituted by Configurations 121 and 212. These are *impossible patterns*! The first shows a decrease in aggressive impulses that is followed by an increase, complemented by an inversely U-shaped quadratic trend. This pattern is contradictory. Therefore, validly, the program did not find anyone showing this pattern. This applies accordingly to Configuration 212 which describes an increase, followed by a decrease, complemented by a U-shaped quadratic trend. Figure 20.2 shows the temporal curves for the six possible patterns.

Table 20.4 First-order CFA of first and second differences of 3 waves of data on aggressive impulses

Configuration[a]	m_{ijk}	$\hat{m}_{ijk}$	z	p	*Type/Antitype?*
111	20	22.241	−.475	.3173	
112	13	13.470	−.128	.4491	
121	0	14.505	−3.809	.0001	Antitype
122	26	8.785	5.808	.0000	Type
211	36	20.733	3.353	.0003	Type
212	0	12.557	−3.544	.0002	Antitype
221	15	13.521	.402	.3438	
222	4	8.189	−1.464	.0716	

[a] The order of variables is DAI1, DAI2, and DAI2ND

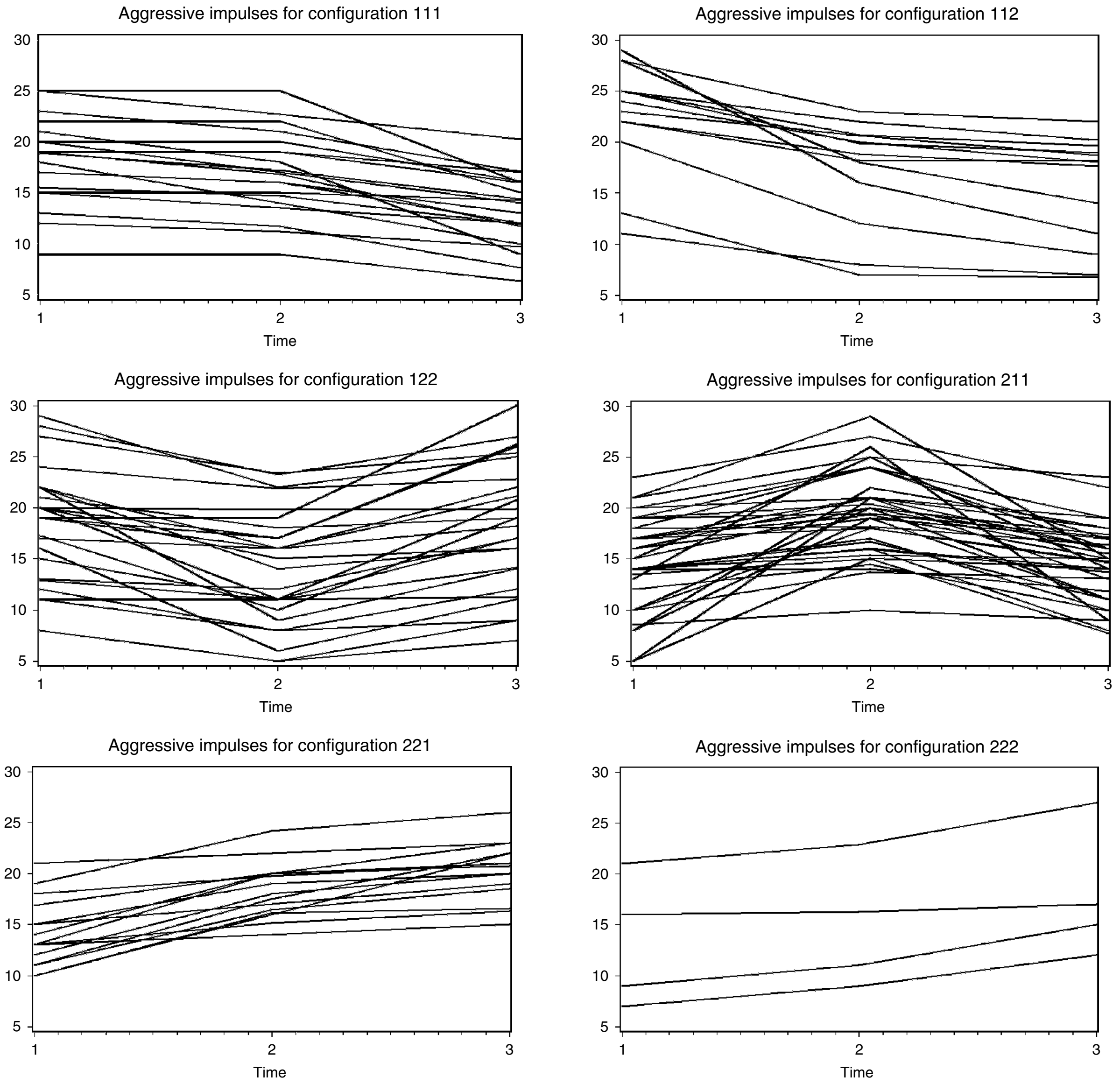


Figure 20.2 Observed aggressive impulse scores for six longitudinal configuration trends

Now, the inspection of Table 20.4 shows that the base model estimated expected frequencies for the two impossible patterns. Only because of these expected frequencies, can these two configurations be said to constitute antitypes. It is, however, a mistake to estimate expected frequencies for cells that cannot contain any case, i.e., for cells with *structural zeros*. We therefore recalculate the CFA, taking the two structural zeros into account. The results are summarized in Table 20.5. Please note that, because of the two structural zeros, the Bonferroni-adjusted α now is 0.008 instead of 0.006 as in Table 20.4.

The log-linear base model for the results in Table 20.3 came with an overall goodness-of-fit LR-$X^2 = 88.36$ ($df = 4$; $p < 0.01$). The log-linear base model for the results in Table 20.4 comes with an overall goodness-of-fit LR-$X^2 = 25.39$ ($df = 2$; $p < 0.01$). The difference between these nested models, $\Delta X^2 = 62.97$, is significant ($\Delta\ df = 2$; $p < 0.01$). This result indicates that taking into account the structural zeros improves the model considerably. However, the base model is still rejected. Therefore, we can expect types and antitypes to emerge.

The two types from the first analysis still exist. They may be less pronounced as in Table 20.4 (as seen by the decrease in the size of the raw residual), but they are still strong enough to be significant even when the conservative Bonferroni procedure is applied.

Naturally, the next question is whether the non-empirical element in the Configurations 122 and 211 prevents them from being interesting (for a discussion of non-empirical or a priori elements in empiricial research, see Brandtstädter, 1982; Smedslund, 1984). The answer to this question is no. CFA does not answer the question as to whether configurations are possible (Configurations 122 and 211 are possible). Instead, CFA asks whether configurations are observed at a different rate than expected. The information about the quadratic trend is, in these two configurations, not informative. Therefore, appending the "1" to the "12" and the "2" to the "21" does not alter the picture. The types therefore indicate that the trends "12" and "21" were observed more often than expected. However, a "1" or a "2" does carry additional information for the Configurations 11. and 22., where the period indicates that either of the shapes of the quadratic trend could be considered.

We conclude that the present analysis contains three groups of cells. In the first, all variables carry information. These are cells 111, 112, 221, and 222. In the second group of cells, only two of the three variables carry information. These are cells 122 and 211. The cells in the first two groups are possible in the sense that these patterns can be observed. The third group of cells contains impossible patterns. This group comprises Cells 121 and 212. In the analysis, these cells have to be declared structural zeros.

3.3 CFA under consideration of a priori probabilities

By far the majority of CFA applications estimates expected cell frequencies from the data.

Table 20.5 First-order CFA of first and second differences of 3 waves of data on aggressive impulses, structural zeros taken into account

Configuration[a]	m_{ijk}	$\hat{m}_{ijk}$	z	p	*Type/Antitype?*
111	20	27.886	−1.493	.0854	
112	13	17.234	−1.020	.0783	
121	0	0	–	–	
122	26	13.880	3.253	.0006	Type
211	36	23.880	2.480	.0066	Type
212	0	0	–	–	
221	15	19.234	−.965	.1672	
222	4	11.886	−2.287	.0111	

[a] The order of variables is DAI1, DAI2, and DAI2ND

However, there exist situations in which determining expected cell frequencies from a priori probabilities is an option. CFA of differences is such a situation. For first differences, we illustrated this using the data situation in the last example. Consider a dataset from three points in time, where all data points are different (for the situation in which data points can be unchanged, see von Eye, 2002). Let the scores be 1, 2, and 3. For these scores, the possible sequences are given in Table 20.6, along with the resulting sign patterns, their frequencies and probabilities.

Clearly, patterns $+-$ and $-+$ come with a probability that is twice that of the probabilities of patterns $++$ and $--$. In addition, these differences are not a matter of data characteristics. They are a priori and will be found in every dataset with three or more observation points. Researchers may wish to take the different a priori probabilities into account. Let the probability of the jth sign pattern of the ith variable be π_{ij}. Then, the expected cell frequency for this transition pattern is estimated to be $e_{ij} = \pi_{ij}\, N$. If the transition patterns for two variables are crossed, the expected cell frequencies can be estimated as $e_{(ixk)j} = \pi_{ij}\, \pi_{kj'}\, N$, where j indicates the sign pattern of variable i and j' indicates the sign pattern of variable k.

To illustrate, we reanalyze the data in Table 20.4, taking into account the a priori probabilities of the four different transition patterns. Table 20.7 shows the results.

The cell probabilities in Table 20.7 were calculated in three steps. First, the a priori probabilities were determined as described above. The same probabilities result as given in Table 20.6. Second, taking into account the specific data situation in Table 20.5 (structural zeros), the probabilities for negative second differences were weighted with 59/114, and the probabilities with positive second differences were weighted with 55/114, to reflect the marginal probabilities of the two signs of the second differences. Finally, the structural zeros were blanked out, and, therefore, the probabilities of patterns 122 and 211 were multiplied by 2.

To give an example, the a priori probability of Configuration 111 is 1/6. This is weighted with 59/114 which yields $\hat{p}_{111} = 0.16667 \cdot 59/114 =$ 0.0863. Multiplied by $N = 114$, we obtain the estimated expected cell frequency given in Table 20.7.

The results in Table 20.7 show that there are no types and antitypes when a priori probabilities are taken into account. It is a typical result that taking different information into account when estimating expected cell frequencies leads to different type and antitype patterns. Currently, the user has to make a decision as to whether to consider data characteristics

Table 20.6 Sequences of differences from the scores 1, 2, and 3, their sign patterns, frequencies, and probabilities

Sequences	*Differences between adjacent scores*	*Sign pattern*	*Frequency of pattern*	*Probability of pattern*
123	−1, −1	−−	1	.167
132	−2, 1	−+	2	.333
213	1, −2	+−	2	.333
231	−1, 2	−+		
312	2, −1	+−		
321	1, 1	++	1	.167

Table 20.7 First-order CFA of first and second differences of 3 waves of data on aggressive impulses, structural zeros and a priori probabilities taken into account

Configuration[a]	m_{ijk}	*cell prob.*	$\hat{m}_{ijk}$	*z*	*p*	*Type/Antitype?*
111	20	.0863	8.16	3.242	.0006	
112	13	.0804	10.83	1.266	.1027	
121	0	0	0	–	–	
122	26	.1608	18.33	1.791	.0367	
211	36	.1725	19.67	3.683	.0001	
212	0	0	0	–	–	
221	15	.0863	9.83	1.648	.0497	
222	4	.0804	9.17	−1.706	.0440	

or a priori probabilities when estimating cell probabilities. The currently available software does not allow one to take into account both.

3.4 CFA of transitions from one point in time to the next

When categorical variables are observed over time, the cells in the cross-classifications of the observed variables indicate exactly where an individual came from and where the individual went. For example, Pattern 13 indicates that an individual endorsed Category 1 at the first and Category 3 at the second point in time. Here, we consider two CFA base models that represent two sets of hypotheses.

First-order CFA of transitions between two points in time

The first model to be considered is the model of first-order CFA, i.e., the main effect model. As was indicated above, this is a log-linear main effect model. For the case of the repeatedly observed variable A, this is the model

$$\log m = \lambda + \lambda_i^A + \lambda_j^A$$

where *i* indicate the rows (categories endorsed at Time 1) and *j* indicates the columns (categories endorsed at Time 2) of the cross-classification. Alternatively, the expected cell frequencies can also be estimated using the well known X^2 formula

$$e_{ij} = \frac{m_{i.}m_{.j}}{N}$$

where $_{i.}$ and $_{.j}$ indicate the row sums and the column sums. It should be noted that, for the present kind of data analysis, it is not required that one variable is observed two or more times. The present variant of CFA can be performed even if two different variables are observed at the two points in time. To make this section easier to compare with the following, we use the example of one repeatedly observed variable. In either case, resulting types (and antitypes) indicate that certain transitions are more (or less) likely than expected using the base model of independence between Time 1 and Time 2 observations.

When one variable is observed repeatedly, one would expect types to emerge for the main diagonal, and antitypes for the off-diagonal cells. This result would correspond to a strong autocorrelation. However, in many applications, not all diagonal cells turn out to be types, and not all off-diagonal cells turn out to be antitypes. In different words, CFA allows one to determine which categories carry an autocorrelation (if it exists at all).

Data example

In the following example, we use the vacation data that were analyzed by von Eye and Niedermeier (1999). At Time 1, a sample of 89 children was asked where their families had spent their last vacations. Responses indicating that the families had spent their vacations at a beach, at an amusement park, or in the mountains were included in the analysis. At Time 2, the same children were asked where they would like to spend their next family vacations. Again, beach, amusement park, and mountains were the options. Using first-order CFA, we now answer the question as to whether certain transition patterns are particularly likely or particularly unlikely. To obtain the results in Table 20.8, we used the base model of variable independence, the z-test and the Bonferroni-adjusted $\alpha^* = 0.0056$.

Table 20.8 shows that vacation preferences in children are stable. Each of the diagonal cells (Cells 11, 22, and 33) constitutes a type, indicating that children typically prefer repeating a vacation over trying something new. Of the off-diagonal cells, only one constitutes an antitype, Cell 13. This cell indicates that it is rather unlikely that children who spent their last vacations at the beach will opt for the mountains as their next vacation place. Each of the off-diagonal cells was observed less often than expected. However, only pattern 13 constituted an antitype.

4 CFA of symmetry patterns

The analysis that was done to create the results on Table 20.8 was a routine first-order CFA. We now ask a different question. We ask whether specific transition patterns are more (or less) likely to be observed than their counterpart transitions in the opposite directions. To answer this question, we use the concept of *axial symmetry* (Lawal, 1993, 2001; von Eye and Spiel, 1996). Let p_{ij} be the probability of Cell ij in a square contingency table. Then, this table is said to display axial symmetry if $p_{ij} = p_{ji}$, for $i, j = 1, \ldots, I$. To test the hypothesis that a table displays axial symmetry, the expected cell frequencies are determined in two steps (note that this model is not log-linear; therefore, standard categorical data modeling software can be used only if it allows for user-specified vectors of the design matrix). First, the cells in the main diagonal are blanked out. They do not belong to the cells involved in the hypothesis. Second, for

Table 20.8 First-order CFA of preferred vacations, observed at two points in time

Configuration[a]	m_{ijk}	$\hat{m}_{ijk}$	z	p	*Type/Antitype?*
11	25	13.303	3.207	.0007	Type
12	10	13.303	−.906	.1826	
13	2	10.393	−2.603	.0046	Antitype
21	3	8.270	−1.832	.0334	
22	19	8.270	3.731	.0001	Type
23	1	6.461	−2.148	.0158	
31	4	10.427	−1.990	.0233	
32	3	10.427	−2.300	.0107	
33	22	8.146	4.854	.0000	Type

Table 20.9 First-order CFA of preferred vacations, observed at two points in time

Configuration[a]	m_{ijk}	$\hat{m}_{ijk}$	*z*	*p*	*Type/Antitype?*
11	25	–			
12	10	6.5	1.88	.170	
13	2	3.0	.33	.564	
21	3	6.5	1.89	.170	
22	19	–			
23	1	2.0	.50	.480	
31	4	3.0	.33	.564	
32	3	2.0	.50	.480	
33	22	–			

the off-diagonal cells, the expected frequencies are estimated, for each pair *ij*, as

$$e_{ij} = e_{ji} = \frac{m_{ij} + m_{ji}}{2}$$

This estimation has the effect that the pairs of cells that are mirrored at the main diagonal contain the same estimated expected cell frequencies. These are the configurations that indicate transitions in the opposite directions.

For use in CFA, estimation proceeds as described. The CFA tests proceed as usual. Only the protection of α differs from the usual procedure. The number of tests is smaller than in standard applications of CFA. The number of symmetry pairs in a square table is $\binom{I}{2} = \frac{I(I-1)}{2}$ where I is the number of rows and columns in the table. This number is smaller than I^2. The protection of α is thus less restrictive than in standard CFA. Table 20.9 displays the reanalysis of the data from Table 20.8 under the hypothesis of axial symmetry. We use the X^2-test and the Bonferroni-adjusted $\alpha^* = 0.05/3 = 0.0167$.

Table 20.9 shows that there is not a single symmetry pair that violates the symmetry assumptions. We conclude that the probability of switching from vacation preference *i* to *j* is the same as the probability of switching from *j* to *i*. As was shown in the previous section, the activity in this matrix can be described using the method of first-order CFA.

5 Discussion

In this chapter, we described a small number of the possibilities that configural frequency analysis (CFA) offers for the analysis of longitudinal data. Specifically, we looked at the analysis of transition patterns, trends, and symmetry. We also discussed the comparison of trends across two populations. These and all other configural approaches to the analysis of cross-classification share the person-oriented perspective (Bergman and Magnusson, 1997; von Eye and Bergman, 2003). Under this perspective, data are analyzed with the goal of making statements about people. The currently dominating perspective leads to statements about variables and their relationships. For example, growth curve modeling (Bollen and Curran, 2006; see also Stoolmiller, Chapter 32, in this volume) aim at describing relationships among variables. In contrast, configural frequency analysis aims at detecting salient patterns of change in people.

Therefore, the types and antitypes that result from configural analysis play a different role than the residuals in models of categorical data. Large residuals indicate where a particular model does not describe the data well. Models are usually expressed in terms of relations among variables. Thus, large residuals indicate which of the proposed variable relationships fail to describe the data. The implication of large residuals is that the model must either be rejected or modified such that the correspondence with the data becomes closer. In contrast, types and antitypes contradict a base model that is of substantive interest only because it takes those variable relationships into account that are not important. If types and antitypes emerge, the relationships that are deemed important must exist. However, results are not expressed in terms of these relationships. Instead, results are expressed in terms of the profiles (configurations) of those individuals who were observed at rates contradicting the base model. In addition, instead of altering the base model, researchers then attempt an interpretation of the types and antitypes.

This procedure will not prevent researchers from employing various CFA base models. However, again, using several CFA base models on the same data does not mean that a base model is dismissed. Rather, it implies that the data are approached with different questions. For example, instead of performing a two-group analysis in our first data example, we could have performed a standard first-order CFA. This alternative method of analysis may have led to types and antitypes also. However, these types would have to be interpreted individually. A gender comparison would not have been possible.

Many more methods of longitudinal CFA have been discussed. These methods involve, for example, creating different kinds of differences, the analysis of pre-post designs, CFA of changes in location, CFA of both trends and shifts in means, CFA of series that differ in length, CFA of control group designs, and the analysis of correlational patterns over time (for a description of these and other methods of longitudinal CFA, see von Eye, 2002). The large array of CFA methods for the analysis of longitudinal data shows that the number of questions that can be answered using CFA, i.e., the number of questions that can be approached from a person-oriented perspective, is equally large.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edn. New York: Wiley.

Bergman, L.R. and Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, 9: 291–319.

Bollen, K. A., and Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective*. New York: Wiley.

Brandtstädter, J. (1982). A priorische Elemente in psychologischen Forschungsprogrammen [a priori elements in psychologische research programs]. *Zeitschrift für Sozialpsychologie*, 13: 267–277.

Finkelstein, J. W., von Eye, A. and Preece, M. A. (1994). The relationship between aggressive behavior and puberty in normal adolescents: A longitudinal study. *Journal of Adolescent Health*, 15: 319–326.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1: 141–149.

Goodman, L. A. (1984). *The Analysis of Cross-Classified Data having Ordered Categories*. Cambridge, MA: Harvard University Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6: 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75: 383–386.

Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika,* 76: 624–625.

Indurkhya, A. and von Eye, A. (2000). The power of tests in configural frequency analysis. *Psychologische Beiträge*, 42: 301–308.

Küchenhoff, H. (1986). A note on a continuity correction for testing in three-dimensional configural frequency analysis. *Biometrical Journal,* 28: 465–468.

Lawal, H.B. (1993). Association, symmetry, and diagonal models for occupational mobility and other similar square contingency tables having ordered categorical variables. *Biometrical Journal*, 35: 193–206.

Lawal, H.B. (2001). Modeling symmetry models in square contingency tables. *Journal of Statistical Computation and Simulation*, 71: 59–83.

Lehmacher, W. (1981). A more powerful simultaneous test procedure in configural frequency analysis. *Biometrical Journal,* 23: 429–436.

Lehmacher, W. (2000). Die Konfigurationsfrequenzanalyse als Komplement des log-linearen Modells [Configural Frequency Analysis complements the log-linear model]. *Psychologische Beiträge*, 42: 418–427.

Lienert, G.A. and Krauth, J. (1975). Configural frequency analysis as a statistical tool for defining types. *Educational and Psychological Measurement*, 35: 231–238.

Perli, H.-G., Hommel, G. and Lehmacher, W. (1985). Sequentially rejective test procedures for detecting outlying cells in one- and two-sample multinomial experiments. *Biometrical Journal,* 27: 885–893.

Rudas, T. (1998). *Odds Ratios in the Analysis of Contingency Tables*. Thousand Oaks: Sage.

Schuster, C. and von Eye, A. (2000). Using log-linear modeling to increase power in two-sample Configural Frequency Analysis. *Psychologische Beiträge*, 42: 273–284.

Smedslund, J. (1984). What is necessarily true in psychology? *Annals of Theoretical Psychology*, 2: 241–272.

Steiger, J.H., Shapiro, A. and Browne, M.W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika,* 50: 253–264.

von Eye, A. (2002). *Configural Frequency Analysis – Methods, Models, and Applications*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A. (2004). Base models for Configural Frequency Analysis. *Psychology Science*, 46: 150–170.

von Eye, A. and Bergman, L.R. (2003). Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. *Development and Psychopathology*, 15: 553–580.

von Eye, A. and Bogat, G.A. (2004). Testing the assumption of multivariate normality. *Psychology Science*, 46: 243–258.

von Eye, A. and Gardiner, J.C. (2004). Locating deviations from multivariate normality. *Understanding Statistics*, 3: 313–331.

von Eye, A. and Gutiérrez-Peña, E. (2004). Configural Frequency Analysis – the search for extreme cells. *Journal of Applied Statistics*, 31: 981–997.

von Eye, A. and Mun, E.Y. (2003). Characteristics of measures for 2 × 2 tables. *Understanding Statistics*, 2: 243–266.

von Eye, A. and Niedermeier, K.E. (1999). *Statistical analysis of longitudinal categorical data – An introduction with computer illustrations*. Mahwah, NJ: Lawrence Erlbaum.

von Eye, A. and Schuster, C. (1998). On the specification of models for Configural Frequency Analysis – Sampling schemes in Prediction CFA. *Methods of Psychological Research – Online*, 3: 55–73.

von Eye, A., Schuster, C. and Gutiérrez-Peña, E. (2000). Configural frequency analysis under retrospective and prospective sampling schemes – frequentist and Bayesian approaches. *Psychologische Beiträge*, 42: 428–447.

von Eye, A. and Spiel, C. (1996). Standard and nonstandard log-linear symmetry models for measuring change in categorical variables. *The American Statistician*, 50: 300–305.

von Weber, S., Lautsch, E. and von Eye, A. (2003a). Table-specific continuity corrections for configural frequency analysis. *Psychology Science*, 45: 355–368.

von Weber, S., Lautsch, E. and von Eye, A. (2003b). On the limits of configural frequency analysis: Analyzing small tables. *Psychology Science*, 45: 339–354.

von Weber, S., von Eye, A. and Lautsch, E. (2004). The type II error of measures for the analysis of 2 x 2 tables. *Understanding Statistics*, 3: 259–282.

Chapter 21

Analysis of longitudinal categorical data using optimal scaling techniques

Catrien C. J. H. Bijleveld

1 Introduction

Optimal scaling techniques can be used flexibly for the analysis of mixed measurement level longitudinal data. After giving some basic technical background, it will be shown and illustrated how, by using a special set up of the data matrix, developmental patterns can be explored, and how various types of growth curves may be modeled. Ending with a real-life example from criminological research, it is shown how solutions can be interpreted and how development for subgroups of respondents can be visualized. One advantage of the techniques discussed is that they can handle missing values and missing occasions flexibly and that they are not burdened by temporal dependence the way ordinary statistical techniques are; a disadvantage is that criteria for interpretation are fuzzy, and that no statistical tests are provided.

This chapter discusses the exploratory analysis of longitudinal data. In doing so, it will focus on techniques that have essentially been developed for the multivariate analysis of mixed measurement level variables. Examples are the analysis of the association between marital status, profession, and income, or between type of therapy, psychological complaints, and therapist attachment. In these examples, at least one variable has been measured at less than interval level, and is noncontinuous: the measurement scale of these variables consists of a set of categories. Categorical variables partition the subjects into different categories, like “marital status” (married, divorced, single, widowed) or “offence” (property, violence, public order, other). In the literature such variables are also referred to as “ordered categorical” and “interval categorical”, or “discrete interval”; see Agresti (1990).

The techniques developed for the analysis of such data that will be discussed here are generally referred to as optimal scaling techniques, although there is more they have in common than their algorithm. Since the first publications on this type of technique, they have become much more easy to use in software packages such as SPSS, and to a lesser extent SAS.

This easy availability of these techniques has not made for their widespread use, however. The optimal scaling techniques, while very flexible and accessible even for a nonstatistical audience, have remained somewhat esoteric and have not yet become as widely used as similarly “novel” techniques, such as multilevel analysis.

Optimal scaling techniques serve essentially as a tool for interpretation, for structuring the data. No distributional assumptions are made.

It is exactly this property which makes optimal scaling techniques so eminently useful for the analysis of patterns of change. The fact that there is temporal dependence in the data and that standard statistical tests cannot be performed, does not hamper these techniques. Because of their flexibility, they make it possible to essentially do-it-yourself model all kinds of growth curves, or dependencies, and assess these models as to their explanatory contribution to theory, or as to exploration itself. Lastly, optimal scaling techniques have efficient options to deal with missing values. Dropout or attrition is a serious problem in many surveys, and is often especially problematic in longitudinal research, as dropout is, with increasing time points incremental, and as data from respondents lost to follow up are virtually worthless if the analysis technique requires a complete series of measurements. As will be shown below, optimal scaling techniques are also very flexible in this respect, and make it possible to make maximum use of the data as they have been collected. For that reason as well, they are thus a commendable option for the analysis of longitudinal data.

The fact that optimal scaling techniques do not make distributional assumptions is on the one hand an advantage, but may on the other hand also be a disadvantage. While fit measures are provided, no statistical test is provided for the goodness of fit of the model. Permutational methods such as bootstrapping or jackknifing may be used to arrive at such a measure, but this is cumbersome. This is probably, in part, the reason that these techniques have not become part of the standard toolkit of the typical (social) scientist. A second reason may be that for interpretation no hard criteria are given such as rotation criteria in factor analysis, although general guidelines and rules of thumb are available.

In the following, we will first briefly discuss the optimal scaling techniques. Such a discussion can be cursory only, and the reader is referred to standard works such as De Leeuw (1983), De Leeuw (1989), Gifi (1990), Van de Geer (1993), Greenacre (1984), Greenacre (1993), Nishisato (1994) and Van der Heijden (1987) for more detail and depth. Next, we will discuss two types of techniques, namely one-set analysis and multiset analysis. Next, we will show how these can be used to investigate longitudinal data in general, and to model growth more particularly. While we give small artificial examples throughout, we end with a real-life example in which we investigate the development of norm-transgressing behavior for three cohorts of secondary school students over three waves. We end with a number of less easily accessible extensions to these techniques. In our descriptions, we at times refer back to Bijleveld and Van den Burg (1998), who presented a more extensive and detailed version of the reasoning in this chapter.

2 Optimal scaling

The various techniques in this chapter have in common that they give different values to existing category values. This process is called quantification or optimal scaling. It is optimal because a certain criterion is optimized. After the variables have been quantified they are treated as if they were continuous.

The optimal scaling techniques use an alternating least squares (ALS) algorithm to arrive at a solution. This is not essential, as the same analyses can be performed using alternating maximum likelihood estimation as well (see De Leeuw, 2006a); also, more sophisticated optimization algorithms have recently been proposed (De Leeuw, 2005; 2006b). ALS algorithms consist of at least two alternating steps. In each step, a loss function is minimized. Alternating between the steps, the algorithm converges to one solution for the parameters of the (linear) analysis model as well as for the new values for the categories of the categorical variables.

We distinguish between nominal, ordinal, and numerical variables. For nominal variables the values of the categories reflect similarity and dissimilarity. When quantifying the categories of these variables, this information must be preserved. For ordinal variables, the values of the categories have no other information content than the partitioning and the ordering. When rescaling, the ordering must thus be preserved, meaning that (weak) monotone transformations are allowed. For numerical variables, essentially interval level variables, only linear transformations are permitted, as in ordinary multivariate analysis techniques.

As stated, two kinds of models can be analyzed using optimal scaling techniques. The first kind of model analyzes the relations between the variables all belonging to one set. Linear analogs of such models are principal components analysis, or factor analysis. We refer to its nonlinear analog as multiple correspondence analysis; we treat its extension, nonlinear principal components analysis, cursorily. The second type analyzes the relations between the variables distinguished into two or more sets. The best example of such a linear model is canonical correlation analysis, and its nonlinear parallel is called—unsurprisingly—nonlinear canonical correlation analysis, or in the case of more than two sets of variables, nonlinear generalized canonical analysis.

2.1 Multiple correspondence analysis

Multiple correspondence analysis investigates the association between various categorical variables simultaneously. For investigating the bivariate association between two nominal variables, cross-tabulations are generally used. If we are investigating M variables, we are then however faced with the task of inspecting $M(M\text{-}1)/2$ cross-tabulations of bivariate relations. However, if we are interested in the multivariate associations amongst the various variables, all such bivariate inspections are tedious and essentially uninformative, and, as we are interested in the multivariate association, we would like to summarize the most pertinent multivariate relations in the data.

Multiple correspondence analysis is a useful technique for doing so. In multiple correspondence analysis, the categories of the nominal variables are quantified in such a way that the correlation or similarity between all quantified variables is maximal. This is the criterion against which the quantification is optimized.

A nominal variable is characterized by its so-called indicator matrix, the matrix of dummy variables that show to which categories a respondent belongs (1) and to which categories he or she does not belong (0). Suppose that we have measured N respondents on M categorical variables. The indicator matrix of variable j with k_j categories is denoted by the matrix $\mathbf{I}j$, which has size $N \times k_j$ (in the Gifi-system $\mathbf{I}_j$ is generally referred to as $\mathbf{G}_j$). The quantifications of the categories are stacked in a vector $\mathbf{b}_j$(of length k_j). Then the expression $\mathbf{I}_j\mathbf{b}_j$ (resulting in a vector of length N) gives the transformed or quantified variable. This matrix product thus contains the k_j quantifications of the categories of the variables.

As said, correspondence needs to be maximized between the variables. Maximum similarity or maximum homogeneity is attained when all quantified variables are—simultaneously—as similar as possible. This is achieved by next introducing a vector of so-called object scores, that contain scores for the N respondents. This vector is referred to as $\mathbf{z}$, and has length N. By maximizing the similarity between the quantified variables and these object scores, the correspondence between the variables is maximized. In formula this is understood as minimizing the difference between $\mathbf{I}_j\mathbf{b}_j$ and $\mathbf{z}$, which means that for achieving maximum similarity the following loss function needs to be minimized:

$$\sigma(\mathbf{z}, \mathbf{b}_1, \ldots, \mathbf{b}_\mathrm{M}) = \sum_{j=1}^{M} SSQ(\mathbf{z}\text{-}\mathbf{I}_j\,\mathbf{b}_j) \qquad (1)$$

where SSQ (.) stands for the sum of squares of the elements of a vector, with elements written in deviation from the column mean, so that SSQ (.) corresponds with the usual sum of squares notation.

Of course, (1) can be solved trivially by setting all elements of $\mathbf{z}$ equal to zero as well as all elements of $\mathbf{b}_j$. Therefore some kind of standardization has to be imposed: in this case it is requested that $\mathbf{z}'\mathbf{z} = N$. Also, the object scores are generally written in deviation from their mean. Solving expression (1) under these conditions gives a solution in which each respondent is assigned one object score, and each category one quantification. As in comparable techniques, such as principal components analysis, higher-dimensional solutions can be reached as well. To do so, for each subject, a point in p-dimensional space is found, and, for each category, p quantifications are found. This is achieved by extending the N-dimensional vector $\mathbf{z}$ to an $(N \times p)$ matrix $\mathbf{Z}$ (with $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$), and by combining the p category quantifications for each category of each variable in M respective $(k_j \times p)$-dimensional matrices $\mathbf{B}_j$. The loss function is extended similarly and optimization otherwise is the same.

The dimensions thus found are not only orthogonal, but they are also nested. By this we mean that the dimensions found are independent of the dimensionality of the solution sought: whether one seeks a solution in 2, 3 or 4 dimensions, the lower dimensions are always identical, independent of any higher dimensions modeled. In other words: the first p dimensions of a $p+1$, $p+2$, $p+3$ etc. solution are stable.

Multiple correspondence analysis is also known as homogeneity analysis (see Gifi, 1990; Greenacre, 1984; Nishisato, 1980; Tenenhaus and Young, 1985; Van de Geer, 1993). In case of only two variables, multiple correspondence analysis is commonly referred to as correspondence analysis (see Benzécri, 1973). Multiple correspondence or homogeneity analysis is available in the computer program HOMALS (short for HOMogeneity analysis through Alternating Least Squares), in the SPSS procedure CATEGORIES (SPSS, 2006). Other programs also perform multiple correspondence analysis, for instance the SAS procedure CORRESP (see SAS, 2006a).

A number of guidelines exists for interpreting solutions. A first measure of these is the loss, or the badness-of-fit, with a goodness-of-fit measure the number of dimensions minus the loss. The total goodness-of-fit can be broken down to a fit measure per dimension, the so-called eigenvalue. Each dimension of the solution has an eigenvalue, which corresponds to the mean variance of all the variables as quantified on that dimension. As such, the eigenvalues reflect explained variance, per dimension, just like they do in techniques like principal components analysis, although here they reflect the variance of the quantified variables. The lower dimensions always explain the most variance, and as one adds dimensions, the eigenvalues of the added dimensions become lower and lower. As such, for choosing the dimensionality of the solution, a scree plot can be used, although in practice two dimensions are often chosen.

Just as the importance of the respective dimensions can be assessed using the eigenvalues, so the importance of the respective variables can be assessed using the so-called discrimination measures. The discrimination measures reflect to what extent the quantified categories of a variable discriminate well between respondents. This translates into saying that the larger the spread of the quantified categories is, the larger the discrimination measure. Conversely, a variable whose categories are very close together in the solution has little contribution to the structuring of respondents in the solution. Discrimination measures for variables are computed per dimension: so a variable in a two-dimensional solution has a separate discrimination measure for each of the two dimensions. The maximum is 1, the minimum is 0. Thus, it may be that a variable discriminates well between respondents on one dimension, as reflected in a high discrimination value, but does not do so on another dimension. As such, discrimination measures are helpful tools in interpreting the various dimensions of the solution: by inspecting what variables "load" heavily on a dimension, and what variables load more on other dimensions, one can interpret dimensions as scales.

Eigenvalues and discrimination measures are thus both indicators of goodness-of-fit: the eigenvalue gives the explained variance of a dimension, and the discrimination measures tell to what extent the respective variables are associated with the variability in object scores. Eigenvalues and discrimination measures are intrinsically linked, with the eigenvalue of a dimension computed as the average of the discrimination measures of all variables for that dimension. Also, the sum of the p-eigenvalues equals the goodness-of-fit of the total solution.

Each respondent is assigned a point in p-dimensional space. These object scores can be represented in a p-dimensional space, forming a cloud of points. They have unit variance for each dimension. The category quantifications are similarly placed in this space: each category receives a p-dimensional quantification, and these quantifications are as such also points in the same p-dimensional space. Just as the discrimination measures and the eigenvalues are related, so is there a relation between the orthonormalized object scores and the category quantifications. More precisely, the quantification of a category is the average of the object scores of all respondents who scored that category. Conversely, the object score of a respondent is the average of all category quantifications of the categories he or she scored.

This implies that respondents are generally placed in the p-dimensional space close to the categories they scored. Thus, we may characterize a respondent by the categories that are placed in his or her vicinity. This implies also that respondents who are placed in each other's proximity in the solution, will have similar answering patterns. Thus, respondents with similar answering patterns form clusters of object scores around or in the neighborhood of the categories that characterize them. This also implies that if all respondents share one characteristic, this characteristic will be placed centrally; by necessity, for every dimension the quantification will then be zero. Such centrally placed quantifications also occur for categories that are not really characteristic for certain (subgroups of) respondents. Thus, the technique produces a solution in which homogeneous subgroups of respondents with particular response patterns are identifiable. Respondents not belonging to such a subgroup, or categories not shared by such distinct groups, end up in the middle of the solution.

Example of multiple correspondence analysis

For illustrative purposes a small dataset will be analyzed. The dataset contains information on nine colleagues of the author as well as the author herself. Data have been collected by observation on "Profession" (variable 1), "Gender" (variable 2), "Hair color" (variable 3), "Hair length" (variable 4) and "Body length" (variable 5). The data are in Table 21.1.

We ran HOMALS in two dimensions for this dataset. The total fit of the solution is 1.003236, which may be considered acceptable. The eigenvalue of the first dimension was .629; that of the second dimension was .374 (thus, the total fit equalled 2 − (.629 + .374)). The discrimination measures of the quantified variables on the two dimensions are given in Table 21.2. Averaging the discrimination measures on the first and second dimensions gives the eigenvalues.

The eigenvalues show that firstly—as was to be expected—"Profession" is not able to discriminate respondents at all. As all respondents were criminologists there is simply no variance and so also there can be no discrimination on this variable. Next, the variables that

Table 21.1 Data in example dataset

Respondents	*Profession*	*Gender*	*Hair color*	*Hair length*	*Body length*
Victor	criminologist	male	blond	medium	tall
Catrien	criminologist	female	blond	long	shortish, but not diminutive
Miriam	criminologist	female	blond	long	average
Gerben	criminologist	male	gray	short	tall
Kim	criminologist	female	red	long	shortish, but not diminutive
Samora	criminologist	female	black	long	shortish, but not diminutive
Jörgen	criminologist	male	black	short	average
Henk	criminologist	male	black	short	shortish, but not diminutive
Michael	criminologist	male	gray	medium	tall

Table 21.2 Discrimination measures and eigenvalues of example HOMALS solution

Variable	*Dimension 1*	*Dimension 2*
Profession	.000	.000
Gender	.836	.092
Hair color	.642	.756
Hair length	.889	.765
Body length	.780	.258
Eigenvalue	.629	.374

discriminate strongest on the first dimension are "Gender", "Hair length" and to a lesser extent "Hair color" and "Body length". In fact, "Gender" discriminates barely on the second dimension. The first dimension thus discriminates respondents on the basis of gender as well as on variables that have a strong association with gender such as "Body length" (men being taller), and "Hair length" (men generally having shorter hair). The second dimension, that has a much smaller eigenvalue, is based mainly on the discrimination of respondents with respect to hair color and hair length. This is probably pretty coincidental here, as out of three respondents with black hair, two had it cut short. Of all women with long hair, two had blond hair. The relations are much more fuzzy, however, as is reflected in the lower eigenvalue.

Table 21.3 gives all category quantifications and the object scores. As can be reconstructed, the centroid of all object scores is zero, as are the centroids of the category quantifications of any variable. Figure 21.1 plots the category quantifications.

Figure 21.2 plots the object scores, and should be viewed as superimposed on Figure 21.1. Combining the information from the two graphs, some nice clustering becomes clear. On the right-hand side, we see all four females. Red hair is placed in this vicinity, exactly on the spot where the only respondent with red hair was placed (1.315, .635). Long hair is placed close by. Blond hair is also close to the category quantification for "female", but as there is also a male with blond hair it tends somewhat to the left-hand side of the picture where most males were placed. The male criminologists are placed in this left-hand side, but much more spread out. The two shortish men with black hair are placed in the bottom left of the picture. Two tallish men with medium-length hair have been placed in the top left-hand side of the figure. There is one man left (Gerben) who does not fit either homogeneous subgroup: this is a tall man (for which reason he should be placed with the upper left cluster) with, however, short hair (for which he should be placed

Table 21.3 Object scores and category quantifications

Object scores			*Body length*			*Gender*		
Victor	−.932	1.176	shortish	.781	−.128	female	1.022	.338
Catrien	.992	.694	average	.264	−.710	male	−.818	−.271
Miriam	.829	.380	tall	−1.217	.645			
Gerben	−1.259	−.222				*Hair length*		
Kim	1.315	.635	*Hair color*			short	−.566	−1.170
Samora	.952	−.355	black	.171	−1.214	medium	−1.196	1.078
Jörgen	−.301	−1.800	blond	.296	.750	long	1.022	.338
Henk	−.138	−1.487	gray	−1.359	.379			
Michael	−1.459	.980	red	1.315	.635	["Criminologist" at (0,0)]		

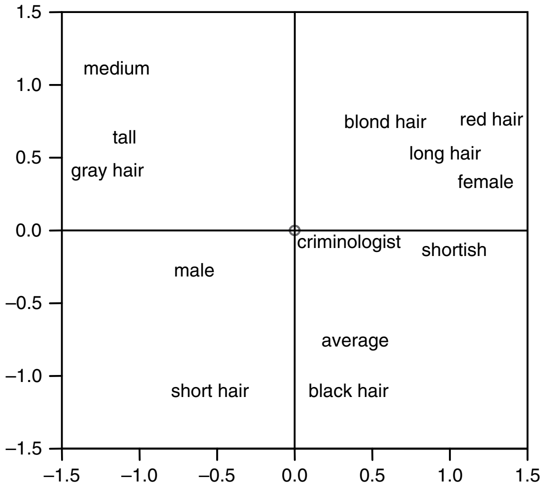


Figure 21.1 Category quantifications of example solution

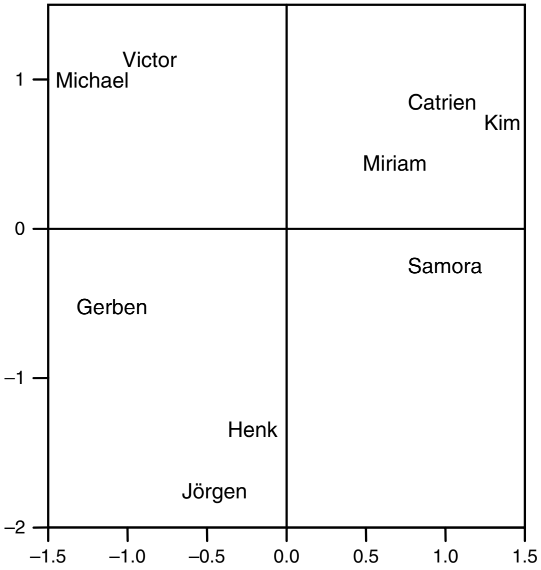


Figure 21.2 Object scores of example multiple correspondence solution

with the lower left cluster). As an intermediate solution, he is placed in between.

For a more in-depth discussion of the HOMALS program, its algorithm, and output, we refer to Gifi (1990), Greenacre (1993), and Van de Geer (1993).

2.2 Extension to ordered and interval variables

In the example above, we treated all variables as categorical variables, i.e., the categories have been treated as if they partition respondents into qualitatively different subsets. The variables "Body length" and "Hair length", however, partition respondents into subsets that reflect a certain ordering as well: from shortish to average to tall, and from short to medium to long. Treating these categories as unordered implies that for these essentially categorically ordered variables we analyze nonlinear relations with the other variables: we look simply at associations at category level. One might however want to investigate relations between variables where this ordering is taken into account. In the context of multiple correspondence analysis this means that one would constrain the category quantifications of the original categories to be on a line, on which they would need to be ordered just like their original ordering. This argument can be extended in the sense that if the categories would reflect not just monotone ordering, but would reflect a certain quantity (say, whether a respondent has 1, 2 or 3 children) one might want these categories to be quantified such that not only the original ordering is preserved, but also the interval properties of the categories. In that case one would constrain the category quantifications to be not only on a straight line, preserving the original monotone ordering, but also request the intervals between the categories to be equidistant.

In case such constraints are desired, we have a set of categorical variables of mixed measurement level: some are nominal, some are ordinal and some are numerical. The technique then to analyse these data, taking into account the mixed measurement level is called nonlinear principal components analysis. Nonlinear principal components analysis can thus be seen as an extension to multiple correspondence analysis, namely where for some variables category quantifications are found under special constraints. Nonlinear principal components analysis can also be viewed as an extension of ordinary principal components analysis, with optimal scaling of categorical variables (Gifi, 1990, Chapter 4; Young, Takane and De Leeuw, 1978).

The first kind of transformation is generally referred to as "nominal", the second as "ordinal", and the third as "interval". In optimal scaling programs the ordinal scaling of category values generally allows also for weak monotone transformations. Also, nominal variables may be transformed such that they are on a straight line, be it that this may be in any order; this is often referred to as "single nominal". In this case, the two-dimensional quantification space has been reduced to one dimension. Recent extensions to the software have also made it possible to opt for more sophisticated transformations of variables such as spline transformations, and have added user-friendly amenities such as imputation of missing values, variable weighting, etc. (see SPSS, 2006). Nonlinear principal components analysis is available in SPSS in the procedure CATEGORIES (SPSS, 2006) and in SAS in the procedure PRINQUAL (SAS, 2006b).

Similarly to multiple correspondence analysis, nonlinear principal components analysis aims at maximizing the similarity between the quantified variables. Again, maximum similarity is attained when all the quantified variables are—simultaneously—as similar as possible, or when all the quantified variables are as similar as possible to the object scores $\mathbf{Z}$, which amounts to minimizing the same loss function as in multiple correspondence analysis, under the condition, however, that the matrices $\mathbf{B}_j$ containing the category quantifications $\mathbf{c}_j$ satisfy

$$\mathbf{B}_j = \mathbf{c}_j\mathbf{a}_j{}'$$

for variables treated as single (i.e., the single nominal variables, the ordinal variables, and the interval variables), where the $p \times 1$ vectors $\mathbf{a}_j$ are the correlations between the transformed variables $\mathbf{I}_j\mathbf{c}_j$ and the object scores $\mathbf{Z}$. These correlations are in the context of optimal scaling referred to as the "component loadings". They are comparable to the discrimination measures in multiple correspondence analysis. Any unrestricted scores in $\mathbf{B}_j$ are referred to as the "multiple category coordinates"; the restricted scores $\mathbf{c}_j\mathbf{a}_j{}'$ are referred to as the "single category coordinates": these are the coordinates of the categories restricted to be on a single straight line in the solution space.

The interpretation of nonlinear principal components solutions runs similar to that of multiple correspondence analysis solutions. Often two dimensions are chosen, and the interpretation is done by looking for clusters of respondents and categories. Again, peripheral patterns of answers and respondents have more particular, characteristic answering patterns and respondents as well as categories placed in the center that do not belong to homogeneous subgroups of respondents with particular profiles of answering patterns. While in multiple correspondence analysis one always uses the category quantifications for interpretations, in nonlinear principal components analysis one does so only for those variables which function in the solution as multiple nominal. For single variables, the single category coordinates are used, i.e., the category quantifications that are restricted to be on a line. Also, whenever single quantifications are present, solutions are not nested.

2.3 Extension to more than one set of variables

In the models discussed thus far, all variables have been treated similarly, i.e., each variable plays the same role in that we investigate the relation of this variable to all other variables simultaneously. However, in many instances it may be the case that we are not interested in the relations between all variables simultaneously, but rather in the relations between one group of variables on the one hand, and another group of variables on the other hand. For example, we might be interested in the relation between a number of socioeconomic characteristics, and a number of demographic characteristics, or, we may be interested in a number of patient personality charateristics, a number of therapist characteristics, and a number of psychodynamic therapy characteristics. In that case, not all variables play the same role. Variables are in such cases conceptually divided into sets, and we are not as much interested in the relations between variables within these sets (e.g., between the personality characteristics); rather, we are interested in the relations between the variables between the sets.

As in the examples above, we are thus looking for homogeneous subgroups of respondents; however, we do not want to investigate their similarity on the basis of the interrelations of all variables, but on the basis of the interrelations of the variables in the different subsets. A linear variety of this technique is referred to as generalized canonical analysis (GCA). When variables of mixed measurement level are included it is called nonlinear generalized canonical analysis (see Van der Burg, De Leeuw and Verdegaal, 1988; and Gifi, 1990, Chapter 5). Nonlinear GCA maximizes the correspondence between two or more sets of variables. For doing so, a weighted sum of the variables in the various sets is constructed and the correspondence between these weighted sums is optimized, just like in canonical correlation analysis or multiple regression analysis. Nonlinear GCA has been implemented in the computer program OVERALS, available in the SPSS procedure CATEGORIES (SPSS, 2006).

The loss function for nonlinear generalized canonical correlation analysis is an extended version of the one for multiple correspondence analysis. The difference is that there are now K sets of variables, each with M_k variables. The indicator matrix of variable j of set k is written as $\mathbf{I}_{jk}$ and the matrix of category quantifications as $\mathbf{B}_{jk}$. The loss function to be minimized is an extended version of loss function (1). Again, the vector $\mathbf{c}_{jk}$ contains the numerical, ordinal or nominal quantification, and $\mathbf{a}_{jk}$ is the vector of weights, now each time for the j-th variable in the k-th set. The badness-of-fit can be broken down by set and by dimension. The eigenvalues again relate to the fit and reflect as before the explained variance of a dimension, although this explained variance now refers back to the weighted sums of variables, and not to the respective variables. This implies that it is theoretically possible to find a good congruence (high fit) between sets, but low explained variance of the variables.

The vector $\mathbf{a}_{jk}$ contains the weights per dimension that sum the variables into the weighted sum, also encountered in the literature under the name canonical variate. As in multiple regression, these weights depend for each variable not only on the particular variable's contribution, but also on the contributions of the other variables within the set. This implies that these weights are susceptible to multicollinearity, so that these can not be interpreted in a straightforward manner. For variables considered as multiple nominal, the scores $\mathbf{B}_{jk}$ are the multiple quantifications of the categories. However, also for multiple nominal variables, interpretation of their relevance is less straightforward than in the one-set case. Analogously to the single variables situation, the p-dimensional quantifications of the categories of multiple nominal variables take into account the effect of the other variables within the set as well: the effect of multicollinearity is incorporated in the multiple category quantifications. Thus, whereas in multiple correspondence analysis the category centroids were identical to the category quantifications $\mathbf{B}_j$, in nonlinear multiset analysis category centroids and category quantifications are different entities. For that reason, for interpreting the association of the categories of a multiple variable, it is advisable to inspect the category centroids of the object scores; for interpreting the association of the categories of a single variable, the projected category centroids must be inspected. Whenever single quantifications are present, nonlinear generalized canonical analysis solutions are not nested, just like nonlinear principal components analysis.

Again, as in the previous two techniques, the object scores again form an orthonormalized system, and category centroids, projected centroids, component loadings, and object scores are part of the same solution and also the same space. Respondents are again characterized by the categories that have been placed in their vicinity, and respondents sharing the same profile form a homogeneous subgroup of respondents; respondents placed at a distance have dissimilar patterns and belong to different subgroups. Categories in the periphery of the solution characterize the more homogeneous subgroups of subjects. Categories in the centre of the solution are shared by many subjects, and generally can not be used to characterize distinct subgroups.

When there are only two sets, nonlinear generalized canonical analysis is the nonlinear analog of canonical correlation analysis (Hotelling, 1936; Tatsuoka, 1988, Chapter 7). In case of two sets of variables and mixed measurement level variables, nonlinear GCA is—but for superficial differences—identical to the model for nonlinear canonical correlation analysis introduced by Young, De Leeuw and Takane (1976) and by Van der Burg and De Leeuw (1983). If there is one variable per set, nonlinear multiset analysis reduces to nonlinear principal components analysis. If there is one multiple nominal variable per set, nonlinear multiset analysis reduces to multiple correspondence analysis. Carroll (1968) defined linear generalized canonical analysis or K-sets analysis; other possibilities for linear K-sets analysis were described by Kettenring (1971) and Van de Geer (1984). For more details on nonlinear multiset analysis, see Van der Burg, De Leeuw and Verdegaal (1988), and Gifi (1990).

3 Analyzing longitudinal data using optimal scaling techniques: two strategies

Optimal scaling techniques can be used flexibly for analyzing longitudinal data. The reasons for this are that, firstly, as these techniques make no distributional assumptions, there is also no need to accommodate the statistical complications that repeated measures induce. Secondly, optimal scaling techniques can fairly flexibly handle missing observations that are an incremental problem in most situations of repeated observations. Thirdly, they have nice features to model and explore growth. Last but not least, categorical variables can be easily included in the analysis.

There are basically two strategies for analyzing mixed measurement longitudinal data in which optimal scaling techniques can be used. In the first, the data box of observations (the first dimension being the respondents, the second dimension being the variables, and the third being the time points) is flattened and analyzed. In the second strategy, some kind of time-series regression is performed. This can either be done by regressing the time-dependent measurements on the time-axis or on some transformation of it, or by accommodating time in some other manner, for instance by regressing the time-dependent measurements on their lagged versions.

If we have measurements over N respondents and k variables, obtained over T time points, we are faced with an $N \times k \times T$ data box. In the first strategy, the data box is first treated as if it were cut into T slices, or matrices, with one $N \times k$ matrix of observations per time point. Ordinary coss-sectional analysis works on this $N \times k$ matrix. The T slices/matrices of observations per measurement wave can be analyzed longitudinally using optimal scaling techniques as follows: a first option is to stack $N \times k$ matrices vertically. Researchers have also employed this method outside of the context of optimal scaling, and the resulting data file is also referred to as a "person-period file", or as a LONG matrix. Each respondent appears as often in the flattened data box as he or she was observed repeatedly. This obviously creates statistical problems if conventional statistical methods are used, because the number of independent observations is less than the NT rows in the data matrix. Thus, statistical tests will be inflated, as the variance of the estimators is underestimated, and the degrees of freedom employed too high. The first problem is not a problem when employing optimal scaling techniques as the techniques serve essentially for exploration. Provisions need to be made to visualize and interpret the development of respondents over time.

In principle, it would also be possible not to flatten the data box by stacking the $T\,N \times k$ matrices horizontally, but vertically (this is referred to as the BROAD matrix). In that case, the $N \times k \times T$ data box becomes an $N \times kT$ data matrix. For many practical situations (as in the previous flattening option, this can also be done when using ordinary analysis techniques) the number of variables may with increasing numbers of time points quickly become too large for analysis. See Visser (1985, pp. 48–55), who formalized these options and discussed statistical implications.

In the second strategy where some kind of time-series regression is performed, the time-dependent measurements are, after flattening the data box to a LONG matrix, regressed on the time-axis or on lagged versions of the measurements. This means that in this strategy we always employ several sets of variables: one that contains the variables of interest and one that contains the regressors. In the context of optimal scaling techniques we will thus almost always use multiset analysis. As an example, we could regress the measurements from time point 1 to time point 10 on a dummy time-variable that has values 1, 2,…, 10. This time-variable can in the context of optimal scaling be treated as an interval variable, in which case we perform some symmetrical version of ordinary time-series regression. We could, however, also treat the time-variable as an ordinal or a spline variable, relaxing the assumptions and not requiring developments over time to be linear but allowing other monotone transformations of change. Of course, we could also allow the time-variable to be single nominal, in which case we freely explore, seeking for maximum correspondence between the variable set and the dummy time-variable set, how the variables develop over time, also allowing for curvilinear or even more irregular development of the variables over time.

Within this second regression strategy we could also regress the variables not on the time-axis itself, but on some lagged version of the variables, in some kind of autoregressive setup. This option is in a sense a variety of the horizontal stacking option (the BROAD matrix) in the flattening strategy. Suppose that we have measurements on two time points. In that case we could simply regress the variables measured at $t = 2$ on the same variables measured at $t = 1$, investigating how the variables at $t = 1$ are associated with the variables at $t = 2$. Or, if we have measured respondents at 4 time points, then we could stack the measurements for $t = 2$, $t = 3$ and $t = 4$ into a first set, and those for $t = 1$, $t = 2$, and $t = 3$ in a second set, such that each respondent's scores at time point $T = t$ in one set are related to his or her scores at time point $T = t{-}1$.

3.1 Examples of the two strategies

We now analyze a fictional example. In the example, 10 respondents have been measured each at 4 time points. For each respondent for each time point we have collected data on the variables "Happiness" and "Sense of purpose". These latter variables are measured on 4-point scales. Respondents are divided into three therapy groups: type A, B and C. See Table 21.4.

Table 21.4 Data in longitudinal example dataset

R	*t*	*Ha*	*Pu*	*Th*[1]	*R*	*t*	*Ha*	*Pu*	*Th*
1	1	1	1	A	6	1	2	2	B
1	2	2	2	A	6	2	2	1	B
1	3	3	2	A	6	3	3	2	B
1	4	4	3	A	6	4	3	3	B
2	1	1	2	A	7	1	1	1	B
2	2	2	2	A	7	2	3	2	B
2	3	3	3	A	7	3	3	3	B
2	4	4	4	A	7	4	4	4	B
3	1	2	2	B	8	1	1	1	B
3	2	2	3	B	8	2	2	1	B
3	3	3	3	B	8	3	2	2	B
3	4	3	4	B	8	4	4	2	B
4	1	2	1	B	9	1	4	2	C
4	2	2	1	B	9	2	3	1	C
4	3	3	1	B	9	3	2	1	C
4	4	3	2	B	9	4	1	1	C
5	1	1	1	A	10	1	4	4	C
5	2	2	2	A	10	2	3	3	C
5	3	3	2	A	10	3	2	1	C
5	4	4	4	A	10	4	1	1	C

[1] R = respondent number, t = time point, Ha = Happiness, Pu = Purpose, Th = type of therapy

We start by performing a one-set analysis, treating each variable as a categorical variable. The data box has been flattened with data matrices for each time point stacked vertically, so that each respondent appears 4 times in the solution: instead of 10 observations the analysis is thus run for 40 observations. We investigate the association between the variables "Happiness", "Purpose", and "Therapy". Doing so, the analysis converges in 26 iterations to an acceptable fit measure of 1.117, with eigenvalues .585 for the first dimension, and .532 for the second dimension. The discrimination measures for the variables per dimension are given in Table 21.5.

As can be seen from Table 21.5, "Therapy" does not discriminate respondents over time points very well: on the first dimension there is hardly any discrimination, on the second dimension it is a little more. This is not surprising as the data show that there are pretty consistent and similar developments over time over the two variables "Happiness" and "Purpose", and the variable "Therapy" is almost orthogonal

Table 21.5 Discrimination measures for the analysis of longitudinal according to flattening strategy

Variable	*Dimension 1*	*Dimension 2*
Happiness	.786	.681
Purpose	.867	.587
Therapy	.103	.327

to these developments. Given that "Happiness" and "Purpose" have such a strong association, it is logical that these two variables dominate the solution over the single "Therapy" variable.

Figure 21.3 gives the structure of category quantifications. It can be seen that high scores on "Happiness" and "Purpose" are located in the right-hand upper part of the figure, and low scores are positioned at the opposite end of the structure. Middle values on these variables are located in the middle bottom part of the figure. This is also where therapy type B is located. Type C is on average located close to low scores of happiness and a sense of purpose. Type A is located somewhat right to the middle, showing that respondents following this therapy are characterized by slightly higher scores on "Happiness" and "Purpose". Respondents following therapy type B tend to be characterized more by somewhat average scores on "Purpose" and "Happiness", and respondents following type C tend to have lower scores on "Purpose" and "Happiness". The so-called horseshoe structure of object scores and category quantifications is typical for multiple

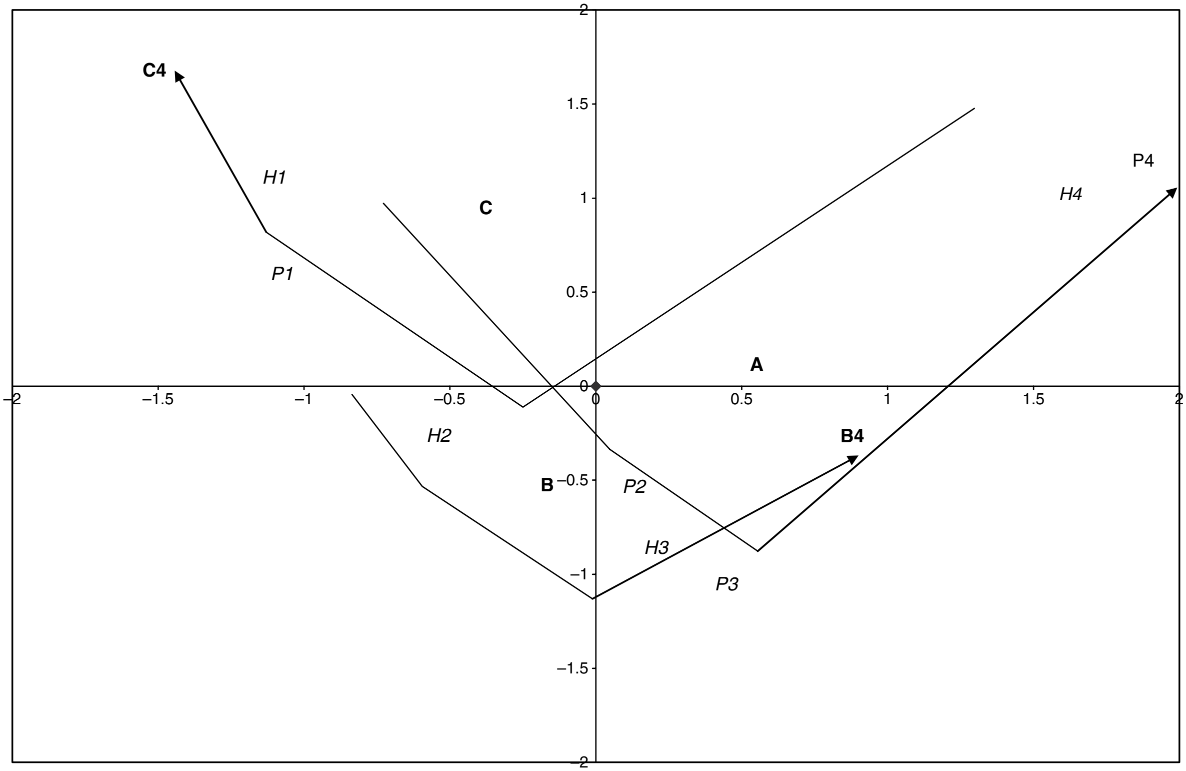


Figure 21.3 Category quantifications and trajectories of therapy groups

correspondence analysis and its associated optimal scaling techniques, and is an artefact of the least squares criterion used.

In Figure 21.3, we have also connected the average object scores of the respondents per therapy group per time point. We could also have connected for each respondent the scores at the various time points, but this gives a not very insightful picture. What the lines now show is how respondents in therapy A in general develop from fairly low scores on “Happiness” and “Purpose” to high scores. This happens with a swinging movement, following the horseshoe. Those in therapy C on the other hand, develop in the other direction; they seem to develop from better to worse. For those in therapy B, developments are positive, though not as marked as for those in therapy A.

Now, we could also perform a regression-type analysis on this data. We will not give the results of such an analysis, as we will be carrying out a regression-type analysis in the empirical example below, but simply outline a number of options for doing this. Firstly, we could simply stack the time-dependent measurements on “Happiness” and “Purpose” in the first set, and construct a dummy-time variable in the second set, which we could treat as either an interval, ordinal, or categorical variable. We could also investigate simultaneously to what extent developments on “Happiness” and “Purpose” are explained by respondents and time points, by adding a second variable to the second set that simply codes the respondents. In this manner we investigate the association between the variables measuring well-being in the first set, and the variables measuring static interindividual differences (the variable indicating the respondent) as well as dynamic intraindividual differences (the time variable) in the first set. See Table 21.6, that gives the variables in the sets, with their measurement levels.

Table 21.6 Design of multiset analysis of therapy outcomes and explanatory static and time-dependent variables

Set 1		*Set 2*	
Respondent	(multiple nominal)	Purpose	(ordinal)
Time	(multiple nominal)	Happiness	(ordinal)

Running such time-regression analyses, one would expect a solution that gives an interpretation not wildly different from the previous analysis. However, there might be some differences, as we now attempt to actually discriminate the time points as well as the respondents on the basis of the variables measuring well-being, while in the previous flattening analysis we simply looked at the association between “Happiness”, “Purpose” and “Therapy”, not taking into account the time dimension.

3.2 The two strategies revisited

In this flattening option in which each respondent appears as often in the data matrix as he or she was observed, it is assumed that changes as they occur materialize in change in the respondents. In other words, respondents change in a stable world. While such an analysis is generally easily carried out and easy to interpret, there are some disadvantages as well. The first of these is spurious effects. If the variables do not interrelate but exhibit the same kind of growth over time, the vertically stacked super-matrix will generate a positive association between the variables. The correlation is spurious because it would disappear when controlling for time. In general, it is thus wise to check solutions of such a vertically stacked matrix also at a number of cross-sections to see whether the associations found over time also exist cross-sectionally and are not merely an artefact of the stacking. Another problem may be that in flattening and vertically stacking the data matrices, the ordering in time of the respondents is lost: if we would permutate the rows of the $NT \times k$ data matrix, we would achieve exactly the same solution. Adachi (2000) provided a solution for this by constraining the time-ordered object scores such that subsequent observations do not receive very different scores, and subsequently (Adachi, 2002) by less restrictively smoothing the object scores over the time dimension and tying them in the time ordering.

In the second flattening option (the BROAD matrix) in which each variable appears as often in the data matrix as it was measured, it is assumed that changes as they occur materialize in change in the variables. The analysis technique views every variable at each time point as a new, different variable. Respondents in other words stay put in a world that evolves around them. This has substantive but also practical implications. If we have, for instance, a variable that indicates the type of drug that respondents use, measured at several time points, the optimal scaling technique regards this variable not as one, but as three variables. Each of these variables is freely quantified, and there is thus no guarantee that the quantifications of the same variable at the various time points will be identical. And if the quantifications are indeed different, the rescaled variable is not comparable anymore across time points, which creates fairly grave conceptual difficulties, and may force the researcher to abandon the analysis. The SERIALS program (Van Buuren, 1990; 1997) ensures that the quantifications of the various temporal versions of the same variable are identical; as this, however, is not implemented in standard software, for practical purposes, the second option must be considered less attractive. In spite of these limitations, it may be a useful option nevertheless, or the only feasible option, when different variables have been measured on subsequent occasions.

Analyzing the LONG or the BROAD matrix has consequences for the handling of missing data. Missing data on a not too large number of variables can be handled fairly easily by optimal scaling techniques. When the $TN \times k$ data matrices are stacked horizontally, respondents lost to follow up can simply be omitted. Suppose that 100 respondents have been observed at time 1, 80 at time 2, and 60 at time 3, the data matrix to be analyzed that results after stacking is simply 240 $(= 100 + 80 + 60) \times k$. All measurement occasions for which data are available can be entered, and those occasions for which no observations were collected are simply left out. This means that series of unequal length can be included, and even interrupted series may be entered. The proportion of missing values should not become too high, however. For instance, in HOMALS the discrimination measures may become larger than one when there are many missing values. Unfortunately, no absolute guidelines on this proportion of missing values can be given.

In the flattening strategy, time as such is thus not explicitly the object of analysis. By stacking respondents horizontally, we are able to explore the “travel” over time of respondents through the p-dimensional space of the solution. In the time-series regression-type strategy, on the other hand, time is much more explicitly modeled, in the sense that the repeatedly measured variables are regressed on variables that are meant to capture the dynamic and time-dependent nature of the measurements. Thus, the time-dependent horizontally stacked variables in one set are related to some version of the time-axis in the other set, and, as we are seeking to maximize the homogeneity between the sets, we are actually attempting not only to summarize respondents’ average scores at each time point, but we are also actually attempting to discriminate the time points in terms of the variables, in order to obtain a profile of each time point in terms of the variables of interest. As such we can answer questions like: What were the general conditions of respondents when they embarked upon therapy? How was their condition midway? and How can their general condition upon completion of therapy be described?

Extending creativity further, of course one could employ not only one time variable and perhaps a variable reflecting interindividual differences, investigating interindividual and intraindividual differences on the variables of interest simultaneously, but we could actually include several dummy time-variables. We could request, for instance, the first variable ($t = 1, 2, 3, \ldots, 11$) to be scaled interval, and incorporate a second variable ($t = 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1$) to be also scaled interval, and as such investigate what growth curve fits the development over time of the variables of interest better. Also, as we usually investigate in two-dimensional space, we could explore whether perhaps on some variables the development is linear or monotone increasing, while on others, as reflected in another dimension, growth is rather curvilinear, or has ups and downs. In this way, we approach hierarchical modeling or growth-curve analysis. See also Michailides and De Leeuw (1997; 2000), who presented a version of multiple correspondence analysis that provides a solution space for all respondents but shrinks or expands the parameters differently for different clusters of respondents.

What we do is not fundamentally different from ordinary time-series regression, apart from the fact that multiset analysis is essentially a symmetrical technique as it looks for association between the weighted sums of the variables of the sets, and does not regress the weighted sum (or in case of only one variable in a set, a variable) on a set of variables. It is possible through some computations to rewrite the results into asymmetrical form (see Bijleveld, 1989, Chapter 4). When there is only one variable in the time-set, it is possible to use the procedure CATREG from SPSS categories, which is a nonlinear analog of ordinary, asymmetrical regression (SPSS, 2006). The ease, again, is that no distributional assumptions are made, as no tests are performed, so the time-dependence can not complicate it either.

Note, that when relating lagged versions of variables, respondents lost to follow up at some later time generate particular problems, as respondents lost to follow up also affect other measurement waves. Also, when relating lagged versions of variables, one essentially looks for what is similar between the measurements at the successive time points; in many situations, the research questions focus more on what changes over time. Of course, it is also possible to regress the variables at $t = 2$ not on the same variables at $t = 1$, but on different variables at $t = 1$. We might, for instance, measure alcohol intake at day t, and headache at day $t + 1$, and see to what extent alcohol intake is associated with headache the next day. Also this analysis could be made fancier by including individual intercepts or separate sets of variables containing categories such as gender, or group membership such as type of therapy. It is also possible to perform such an analysis for one respondent who has been measured on many occasions. Higher-order lags may be included as well.

The analysis of lagged versions of variables, though conceptually attractive, has a number of limitations. Firstly, one loses one time point for every lag: the higher the order of the lags the more time points have to be deleted. Therefore, higher-order lags can become unattractive or even impossible to model when there are few time points. A second technical limitation is that, in rescaling the lagged versions of categorical variables, there is, just like when analyzing the BROAD matrix, no guarantee that the categories of these lagged versions of the variables will receive identical quantifications.

4 Example

We analyze data on self-reported misdemeanours and offending from the Netherlands Schoolproject (Weerman, 2007). During three waves of data collection, several hundreds of students had been interviewed at a number of secondary schools in the urbanized center of the Netherlands. The schools all offered lower-level secondary education, being mainly VMBO (Voorbereidend Middelbaar Beroeps Onderwijs – Preparatory Mid-level Vocational Training) schools, with a number offering tailored vocational training such as an agrarian school and a technical secondary school. In the first wave, students were in either first or third year. In the second wave, students were either second or fourth year (the fourth year being the year in which students in these types of schools do their final exams). In the third wave all students were in their third year. The study design is as such an overlapping cohort design.

At each wave, students were given the same questionnaire that they filled out classwise on the computer, with at least one researcher present to answer questions and help out whenever necessary. The questionnaire asked about a multitude of issues, such as relations with peers, relations with parents, delinquent peers, bonding with school, self-control, and the like. Key dependent variables were delinquency and misdemeanor during the school year. Various demographic characteristics were also registered. The survey is as such a self-report survey. The prominent questions in the Schoolproject were: (1) how does delinquency and misbehavior develop over the course of the school career? and (2) to what static and dynamic predictors are delinquency and misbehavior related?

Not all students were present at all waves. Earlier analyses showed that during the course of the study, there was selective attrition (Weerman, 2006). In our analysis, we were able to include every student at every measurement that was present for him or her. So, if for instance a pupil was present only at waves 1 and 3, these measurements were included in the dataset. If a pupil was present only at the third wave, then only that wave was included. As was explained below, this does not hamper our analysis. Repeat measurements were included as "new cases": the structure of the dataset is in Table 21.7.

Every respondent thus plays a role in the analysis for those waves for which valid measurements were obtained. In total we had thus 4769 rows in the dataset, for a total of 2661 unique students. While missing waves did not hamper the analysis, also missing values did not hamper the analysis as they were simply left missing and did not play a role in the analysis. A number of very sparsely filled categories were recoded to the adjacent category to prevent outliers.

We chose to perform a multiset analysis (OVERALS). Given that the key dependent variables were misbehavior and delinquency, we put these in one set. The delinquency and misbehavior variables, being both variation measures capturing the variation in behavior (with a score of zero implying no delinquent behavior and a high score implying a range of delinquent behavior, which correlates strongly

Table 21.7 Structure example dataset

resp 1	wave 1	variables
resp 1	wave 2	variables
resp 1	wave 3	variables
resp 2	wave 1	variables
resp 2	wave 3	variables
resp 3	wave 1	variables
resp 3	wave 2	variables
resp 3	wave 3	variables
resp 4	wave 2	variables
resp 4	wave 3	variables
resp 5	wave 1	variables
resp 5	wave 2	variables
resp 5	wave 3	variables
resp 6	wave 3	variables
…….		

with the seriousness of delinquency) were both pretty skewed, but as explained above, this doesn't affect our analysis. Both variables were treated as ordinal variables. Next, we put all variables capturing temporal development in a set; these variables were the waves, capturing a period effect (entered twice with a different coding to enable both ordinal as well as nonlinear development). The third variable in this set was the respondent's age, which varied from 11 to 19, with 99% of ages between 12 and 17. The fourth variable was the respondent's grade (which varied from first to fourth year). Both age and grade were treated as multiple nominal variables in order to be able to model nonlinear developments. In a third set we put all static characteristics, these being the school, as well as respondents' sex and ethnicity (coded according to prevailing Statistics Netherlands definitions). A little over 50% were boys, around two-thirds of measurements were obtained for Dutch students, other sizeable ethnic groups were Surinamese (a former colony), Turkish and Moroccan (mainly children of former migrant workers) and other non-western migrants. Antilleans (the Dutch Antilles are still part of the kingdom of the Netherlands) and western migrants were less prominently present. All three variables in this set are nominal variables and were treated as multiple nominal. A last, fourth set was formed by a number of dynamic predictors of delinquency and misbehavior. At each wave, students had indicated whether they had a good relationship with their parents and how strong their bond with school was. These variables figure prominently in theories of juvenile delinquency. Respondents had also indicated whether their friends are delinquent; this variable is a prominent and strong predictor of juvenile delinquency. Lastly, they had rated their own self-control. Most students reported good ties with their parents and average to good attachment to school. Self-control was approximately normally distributed. Seriously delinquent friends were rare. All variables in this set were treated as ordinal. See Table 21.8 for an overview of the sets and their variables.

We ran the analysis in two dimensions and set all analysis parameters otherwise to default. The algorithm converged in to a fit of .86, which may be viewed as acceptable. The object scores were well spread and did not show any indication of outliers. The first dimension had an eigenvalue of .470, the second dimension's eigenvalue was .393. Table 21.9 gives a summary of the loss per set per dimension.

As Table 21.9 shows, the first dimension represents mainly the developments in delinquency and misbehavior, as well as developments in the dynamic predictors over time. The time variables play much less of a role here, as do the—and this is to be expected—static variables. This is not to say that the second dimension is an entirely static one: here the dynamic variables also play a role, although it is less prominent. All in all, it appears—given the loss per set—that delinquency over time is more

Table 21.8 Overview of sets employed in analysis, variables and measurement levels

Set 1		*Set 2*		*Set 3*		*Set 4*	
delinquency	(ordi)[1]	time_1	(ordi)	school	(mnom)	relation parents	(ordi)
misbehavior	(ordi)	time_2	(mnom)	gender	(mnom)	relation school	(ordi)
		grade	(mnom)	ethnicity	(mnom)	self-control	(ordi)
		age	(mnom)	delinquent peers	(ordi)		

[1] mnom = multiple nominal, ordi = ordinal

Table 21.9 Loss per set per dimension

		Dimension 1	*Dimension 2*	*Total loss*
set 1	(delinquency and misbehavior)	.290	.602	.892
set 2	(time variables)	.884	.606	1.490
set 3	(static background variables)	.664	.576	1.240
set 4	(dynamic predictors)	.281	.645	.926
eigenvalue		.470	.393	
fit				.863

strongly related to dynamic predictors that may vary over respondents over time, than to temporal variables that are changing identically over time for each respondent, such as age or grade. Also, quite a bit of interindividual variation is apparently captured by individual differences as measured by the static predictors.

Figure 21.4 gives a representation of the solution. For the multiple nominal variables, the category centroids are depicted. For the ordinal variables the projected centroids are used; the categories of these variables are connected, and they always form a line through the origin, with an arrow pointing to the high scores. For a

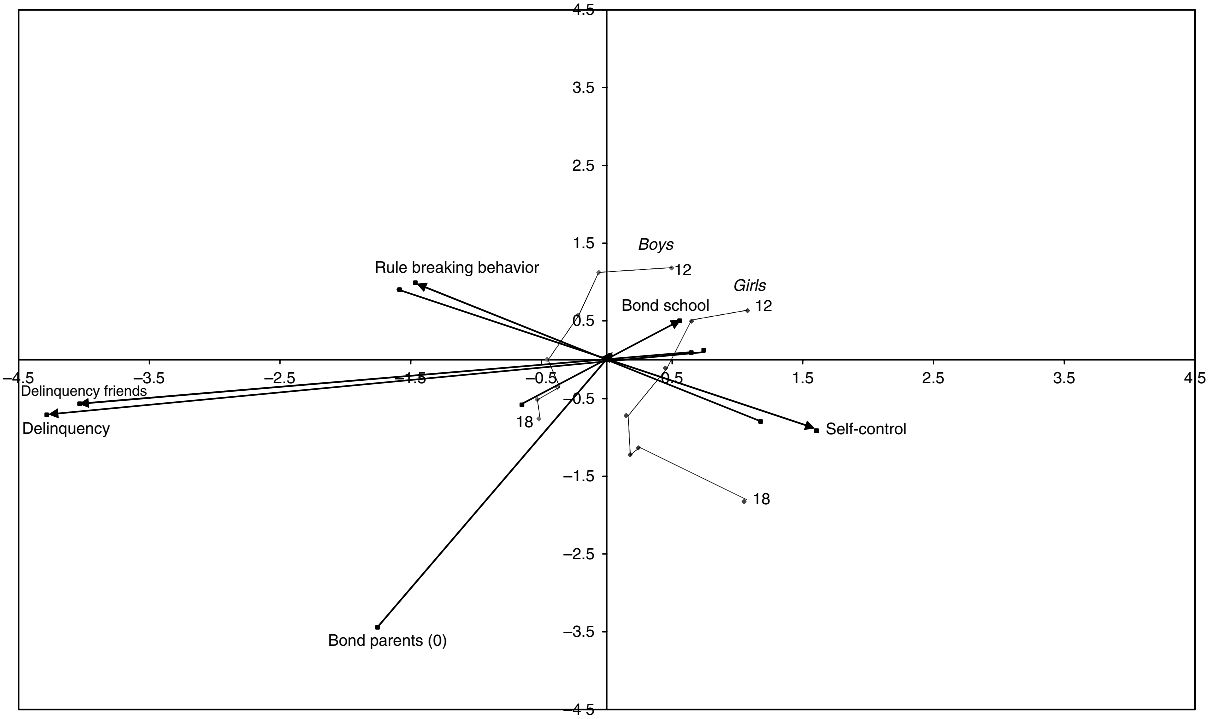


Figure 21.4 Category quantifications and trajectories of boys and girls from age 12 to 18

better overview, not all category centroids have been depicted in the graph.

Interpretation of the solution is done along the lines sketched above; briefly it amounts to investigating the proximity of categories in the solution, and projecting centroids on to those variables whose connected categories form an arrow. Cross-sectional analyses on the datasets revealed approximately the same interpretation, so that we need not be afraid that the relations are spurious. This leads to the following conclusions.

Students who report delinquent behavior at some time point are located in the left-hand side of the figure, somewhat under the x-axis. Students who report misbehavior at some point are located in the upper left side of the figure. It is immediately obvious that delinquency at a certain time point is strongly associated with having delinquent friends. Likewise, misbehaviour is strongly associated with low self-control. The connected categories of the variable measuring students' relationships with their parents at the various waves form an arrow that starts virtually in the center—indicating how almost all students report a good relationship with their parents. It ends in the bottom left of the figure. Students' ties with school at the various waves are also represented as an arrow, that points from a good bond in the middle upper right-hand side of the picture to a not so good bond in the middle lower left-hand side.

Boys (the positions are not indicated in the graph) are positioned on average more in the upper left-hand side of the picture, close to less self-control, and more misbehavior, and also—though less strongly so—more delinquency and more delinquent friends. Girls are placed in the opposite end of the picture, close to less misbehavior and more self-control. The schools are well spread (also not indicated and analyzed here). Antilleans (not indicated, but their average object score is in the left lower side of the picture) stand out for reporting relatively bad relations with their parents, with school, and for relatively high scores on delinquency. Surinamese report about the same level of delinquency but more misbehavior than Dutch. Moroccan and Turkish respondents have a delinquency and misbehavior profile that resembles that of girls most. It is often suspected that especially Moroccan youngsters underreport in self-report delinquency surveys. Given that they are overrepresented in police statistics, we should reckon with the possibility that this is the case here too. Remarkably, hardly any period effects are found; both the ordinal and the multiple nominal version of the cohort variable did not fit notably.

Next, we investigated the development of problem behavior over time by drawing two trajectories. We connected the average object scores of boys and girls at ages 12, 13, 14, etc. up to age 18 (at age 19 we had only a few respondents left). This shows how the trajectories for boys and girls are located respectively in the upper left-hand side and lower right-hand side of the picture. Boys start out at age 12 at an average misbehavior level, a low level of delinquency, and with very good relations with both their parents, as well as school. Self-control assumes an average score for boys at this age, and few report having delinquent peers. Girls start out at low levels of misbehavior, high self-control, ever lower levels of delinquency, and report hardly any delinquent peers; they have good relations with their parents and good ties with school. Thus, at age 12, girls are not delinquent, not misbehaving, while boys are misbehaving a little, but not yet delinquent.

It is remarkable how for boys as well as girls the biggest shift on the first dimension, that captures delinquency and misbehavior, occurs from age 12 to 13. Both swing leftward. For both, this means an increase in misbehavior, and hardly any increase in delinquency. Moving on to age 14, we see how the trajectory for boys and girls starts to move to the bottom half of the picture. For boys this implies that they move to a higher delinquency level and start reporting having delinquent peers. From age 13 onwards, self-control and misbehavior remain fairly constant. For girls, we see basically the same change, though there is a level difference. Both boys and girls now start, from age 14 onwards, reporting less positive feelings towards school. By age 15, delinquency levels are definitely—for boys as well as girls—higher than what they were at age 14; misbehavior levels are more or less the same, and they may even be slightly less by age 15 for boys. Girls' ties with school are now less strong than those of boys, and they also report much poorer relations with their parents than boys do at this age. Self-control remains fairly stable.

This trend continues in similar fashion up to age 16, but by age 17 a remarkable phenomenon takes place as delinquency levels, for boys as well as girls, though more marked for boys, decline. Maybe this has to with the fact that most students at this age are in their final—exam—year. Those students still in the sample at age 18—probably a select group—are marked by increased delinquency levels if they are boys, and by strongly decreased misbehaviour and decreased delinquency if they are girls.

All in all, we see how, over age, misbehavior jumps from ages 12 to 13 and continues to grow only a little and declines towards the end of the school career for boys, a little earlier than for girls. Delinquency increases up to age 16, although the growth levels off after age 15. Relations with school decrease in similar fashion for boys as well as girls; relations with parents decrease much more for girls.

Changes over the various grades could be drawn in the same manner. This would show how for girls the trajectory spreads from the upper part to the lower part of the figure, showing hardly any spread over the first dimension. This shows that over their school career girls are much more stable in their criminal behavior and misbehavior than boys are. Over the school grades, girls in fact develop in the sense of a slight increase in delinquency and a decrease in misbehavior. Boys were shown when the lines were drawn to make a swing in misbehavior, increasing from year 1 to 2, and subsequently—slowly—decreasing. For delinquency the decrease occurs only after year 3.

Not all respondents had filled out the questionnaire at every time point. To investigate whether those respondents who were missing at one or more waves had a different response pattern than those who didn't, we computed the average object score of students with at least one missing wave. It turns out that this average object score is (–.156, –.254), showing that these students are on average more disgruntled with school, report worse relations with their parents, and score relatively high on delinquency and somewhat lower on misbehavior. Those students who were present only on the first wave have an even more marked average location at (–.503, .117), with higher misbehavior and higher delinquency scores. Sample attrition is different for boys and girls: boys who are missing at least one wave have an average object score of (–.442, .071), i.e., high on delinquency and somewhat on the low side on misbehavior, and reporting below average relations with school, but still fairly average relations with their parents. On the other hand, girls who are missing at least one wave have an average object score of (.233, –.696), that implies that these girls score low on misbehavior, report bad relations with school, and particularly so with parents, and score below average on delinquency. Thus, while for boys attrition appears to be associated most with misbehavior and delinquency, for girls bad relations with their parents are characteristic.

5 Extensions

As we showed above, the models discussed here can easily be used to analyze a dataset that has been constructed such that autoregressive types of models can be analyzed. Bijleveld and De Leeuw (1991) presented a model that analyzes the state space model, and allows us to quantify any noninterval variables. The state space model specifies relations between a set of input variables and a set of time-dependent output variables, that are channeled through a latent variable, called the state variable, that captures both the time dependence through an autoregressive model, as well as channels the dependence between input and output variables. The model as such is explicitly asymmetric. The state variable can be one or higher dimensional. While easily applicable, the model has not become used broadly mainly because of lack of user-friendly software. The analysis program DYNAMALS was designed for the analysis of datasets collected on one respondent, that contain many replications over time. DYNAMALS has been extended to also be able to analyze data gathered for more than one respondent. (Bijleveld and Bijleveld, 1997). Quantifications are stable over time and over respondents. The latent state values vary across respondents and over time. For each respondent, a latent state trajectory can be drawn. In general, when higher dimensional solutions are sought, DYNAMALS tends to emphasize interindividual differences in the first dimension(s), and emphasizes intraindividual differences in the following dimensions.

Linear dynamic systems analysis for several subjects with optimal scaling has a number of advantages over ordinary time-series modeling. The most prominent of these is the fact that categorical data can be analyzed. A second type of advantage is in the area of stability. Where cross-sectional methods derive stability from a suitably large number of replications over subjects, and where time-series models derive stability from a suitably large number of replications over time, N > 1 DYNAMALS derives stability from time points as well as respondents.

Van Buuren (1990) developed the SERIALS program that combines a state space or linear dynamic system-type of model with the Box-Tiao transform proposed by Box and Tiao (1977), a method that extracts components from multiple time series, in such a way that these components are related as strongly as possible to lagged versions of themselves. Thus, the correlations between the lag(0) and lag(d) versions of a dataset must be as high as possible. This implies that the technique seeks those components that can be constructed from the data that forecast themselves as well as possible. Because of the forecasting constraint, the extracted components are mostly pretty smooth, and the technique can thus also be viewed as a smoother of wild series. For details, see Box and Tiao (1977) and Van Buuren (1990). The optimal scaling happens under special constraints that ensure that the quantifications of the lagged versions of the variables are identical, so that they can be interpreted as the same variable.

6 Software

Multiple correspondence analysis, nonlinear principal components analysis, and nonlinear multiset analysis are all, probably most easily, available in SPSS in the procedure CATEGORIES. SAS has versions of the first two techniques. Versions in R, often with fancy extensions, are downloadable from Jan de Leeuw's website http://www.cuddyvalley.org/psychoR, which incidentally is a very good website to browse for novel extensions to these techniques.

7 Concluding remarks

In this chapter we discussed a number of exploratory techniques that use optimal scaling to quantify categorical data. The techniques can handle mixed-measurement-level categorical longitudinal data, whose values are rescaled, after which the categorical variables are treated as continuous variables. We discussed two approaches to the longitudinal analysis of categorical data using optimal scaling techniques. In the first, we adapt the data box in such a way that the data can be fed into the technique and analyzed as if they are cross-sectional.

Any categorical variables are transformed. When analyzing the LONG matrix or person-period file, these transformations are stable over time, and it is therefore assumed that the variables themselves do not change over time. When the technique produces a solution we return to the longitudinal properties of the data box, retrieving time points and respondents, and making visible any changes by drawing respondents' trajectories from one time point to the next. Even though this approach is not without methodological hazards—the most notable being the risk of relationships induced by heterogeneity over time points—it is flexible, easy to apply and conceptually attractive. The multiset techniques can be adapted to investigate summary as well as group-specific changes in time. They can be used to explore nonlinear growth for interval-level data, and can incorporate specific additional research questions, such as questions about the discrimination between respondents or groups of respondents. Autoregressive setups can be made. For analyzing mixed-measurement-level longitudinal data, these techniques thus make it possible to investigate flexibly, and untroubled by statistical complications because of serial dependence, various kinds of substantive issues regarding growth.

Another advantage of the use of optimal scaling techniques for the analysis of longitudinal data is that they can handle a modest proportion of missing data relatively easily. The techniques HOMALS, PRINCALS, and OVERALS simply exclude any missing observations from the loss function. The less user-accessible programs for the analysis of long series of data use similar methods. Compared with confirmatory techniques, the two approaches share the disadvantage that no stability of information is provided. For those cases where we have sufficient replications over subjects, this can of course be overcome by jackknifing or bootstrapping, although this is cumbersome.

All in all, the optimal scaling techniques presented here are essentially exploratory techniques useful for generating rather than testing hypotheses. They are relatively untroubled by technical problems due to dropout and attrition common in longitudinal research. They provide a flexible and conceptually attractive framework for investigating a variety of exploratory research questions of a longitudinal nature, incorporating categorical variables of mixed measurement level.

References

Adachi, K. (2000). Optimal scaling of a longitudinal choice variable with time-varying representation of individuals. *British Journal of Mathematical and Statistical Psychology*, 53: 233–253.

Adachi, K. (2002). Optimal quantification of a longitudinal indicator matrix: Homogeneity and smoothness analysis. *Journal of Classification*, 19: 215–248.

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Benzécri, J. P. (1973). L'analyse des données. Paris: Dunod.

Bijleveld, C. C. J. H. and Bijleveld, F. D. (1997). A comparison of the applicability of two approaches to linear dynamic systems analysis for *N* subjects. *Kwantitatieve Methoden*, 55: 37–56.

Bijleveld, C. C. J. H. and De Leeuw, J. (1991). Fitting longitudinal reduced rank regression models by alternating least squares. *Psychometrika*, 56: 433–447.

Bijleveld, C. C. J. H. and Van der Burg, E. (1998). Analysis of longitudinal categorical data using optimal scaling techniques. In Bijleveld, C. C. J. H., Van der Kamp, L. J. Th., *et al. Longitudinal Data Analysis. Designs, Models and Methods*, pp. 46–154. London: Sage.

Box, G. E. P. and Tiao, G. C. (1977). A canonical analysis of multiple time series. *Biometrika*, 64: 355–365.

Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. In Proceedings of the 76th Convention of the American Psychological Association, Vol 5, pp. 227–228.

De Leeuw, J. (1983). The Gifi system of nonlinear multivariate analysis. In E. Diday (ed.), *Data Analysis and Informatics IV*, pp. 415–424. Amsterdam: North-Holland.

De Leeuw, J. (1989). Multivariate analysis with linearization of the regressions. *Psychometrika*, 53: 437–454.

De Leeuw, J. (2005). *Nonlinear Principal Component Analysis and Related Techniques*. UCLA: Statistics Preprints # 427.

De Leeuw, J. (2006a). Principal Component Analysis of Binary Data by Iterated Singular Value. *Decomposition. Computational Statistics and Data Analysis*, 50: 21–39.

De Leeuw, J. (2006b). *Pseudo-Voronoi Diagrams for Multicategory Exponential Representations*. UCLA: Statistics Preprints # 463.

Gifi, A. (1990). *Nonlinear Multivariate Data Analysis*. Chichester: Wiley.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. New York: Academic Press.

Greenacre, M. (1993). *Correspondence Analysis in Practice*. London: Academic Press.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28: 321–327.

Kettenring, J. R. (1971). Canonical correlation of several sets of variables. *Biometrika*, 56: 433–451.

Michailides, G. and De Leeuw, J. (1997). *A Regression Model for Multilevel Homogeneity Analysis*. UCLA: Statistics Series, # 212.

Michailides, G. and De Leeuw, J. (2000). Multilevel homogeneity analysis with differential weighting. *Computational Statistics and Data Analysis*, 32: 411–442.

Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

SAS (2006a). The Corresp procedure. http://v8doc.sas.com/sashtml/stat/chap24/index.htm

SAS (2006b). The Prinqual procedure. http://v8doc.sas.com/sashtml/stat/chap53/index.htm

SPSS (2006). SPSS Categories. http://www.spss.com/categories.

Tatsuoka, M. M. (1988). *Multivariate Analysis: Techniques for Educational and Psychological Research*. New York: Macmillan.

Tenenhaus, M. and Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50: 91–119.

Van Buuren, S. (1990). *Optimal Scaling of Time Series*. Leiden: DSWO Press.

Van Buuren, S. (1997). Optimal transformations for categorical autoregressive time series. *Statistica Neerlandica*, 51: 90–106.

Van de Geer, J. P. (1984). Linear relations between *k* sets of variables. *Psychometrika*, 49: 79–94.

Van de Geer, J. P. (1993). *Multivariate Analysis of Categorical Data: Theory and Applications*. London: Sage.

Van der Burg, E. and De Leeuw, J. (1988). Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36: 54–80.

Van der Burg, E., De Leeuw, J. and Verdegaal, R. (1988). Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53: 177–197.

Van der Heijden, P. G. M. (1987). *Correspondence Analysis of Longitudinal Categorical Data*. Leiden: DSWO Press.

Visser, R. A. (1985). *Analysis of Longitudinal Data in Behavioural and Social Research*. Leiden: DSWO Press.

Weerman, F. (2007, in press). Dutch research on Juvenile Offending. In: M. Tonry and C. Bijleveld (eds), *Crime and Justice in the Netherlands*: *Crime and Justice: A review of Research, Vol 35*. Chicago: University of Chicago Press.

Young, F. W., Takane, Y. and De Leeuw, J. (1978). The principal components of mixed measurement multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41: 505–529.

Young, F. W., De Leeuw, J. and Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 43: 279–281.

Chapter 22

An introduction to latent class analysis

C. Mitchell Dayton

Latent class analysis (LCA) is a method for analyzing categorical data from sources such as achievement test items, rating scales, attitude items, etc. It is assumed that the population from which respondents arose is divided into subgroups within which responses to the variables are independent. These subgroups are not directly observed so the focus of LCA is on characterizing the latent structure of observed categorical data.

1 Introduction

Latent class analysis is a relatively new approach to analyzing multivariate categorical data. Often, for multivariate data the focus of analysis is to explain the interrelations among the variables. Factor analysis, for example, was developed over one hundred years ago (Spearman, 1904) to "explain" the positive intercorrelations among achievement measures. Whereas for continuous variables interrelations are usually quantified by means of correlation coefficients, for categorical variables interrelations involve conditional probabilities that cannot easily be summarized in simple numerical indices. LCA, in its various applications, is based on the notion of conditional independence. Roughly speaking, it is assumed that the population of interest can be divided into non-overlapping subsets of respondents (i.e., latent classes) such that, within each subset, the categorical variables are independent. Note that these subsets are not observed and, indeed, may not even be observable. Consider, for example, variables A and B that have been observed for a sample of respondents. Dependence between A and B would mean that the rates of occurrence of the various categories of B are not the same across the various categories of A. In particular, for two Yes/No attitude items, dependence would mean that the Yes response to B was produced at different rates for respondents saying Yes versus No to A. Such lack of independence is more or less universally observed for categorical variables and a variety of statistical methods has been developed to explore this dependence (e.g., hierarchical log-linear modeling and correspondence analysis). The analytical approach taken in LCA is conceptually similar to factor analysis in that it is assumed that, if latent class membership were known, then the variables would be (conditionally) independent. Note that in classical factor analysis it is assumed that partialling out the effects of the hypothetical (latent) factors reduces the intercorrelations among the variables to zero (i.e., makes them conditionally independent). References that provide historical and theoretical background for LCA include Lazarsfeld and Henry (1968), Goodman (1974), Haberman (1979), Bartholomew (1987), Hagenaars (1990), Von Eye and Clogg (1994), Heinen

(1996), Rost and Langeheine (1997), Dayton (1999), and Hagenaars and McCutcheon (2002).

From a mathematical point of view, LCA is a method for finite mixture modeling where the latent classes represent the components of the mixture. Since the variables are categorical and, in general, no ordering properties for the categories are assumed, each component of the mixture is a product of multinomial probability functions. As mixture models have become more popular in the behavioral sciences, one finds references to latent classes in a broader range of contexts such as item response theory and structural equation modeling, but these applications are not pursued in this chapter.

Before turning to the mathematical model for LCA and the necessary rubrics for estimation, model fit, etc., we present a few types of applications for LCA to provide some context for these methods.

1. Linear hierarchies: the notion of linear hierarchies arises in several research areas including Guttman scaling, developmental sequences, and learning/acquisition sequences (Dayton, 1999). Latent class methods have been applied in this area since the work by Proctor (1970), Goodman (1975), and Dayton and Macready (1976; 1980).
2. Medical diagnosis: latent class methods have been used to assess the value of laboratory tests as diagnostic indicators when no gold standard exists (e.g., Rindskopf and Rindskopf, 1986).
3. Identifying typologies: latent classes may be interpreted as representing clusters of similar respondents within a population and, as such, LCA may be viewed as a modern approach to cluster analysis (Vermunt and Magidson, 2000).

2 Model for LCA

For the ith respondent, let $Y_i = \{y_{ij}\}$ be responses to j = 1,...,J categorical variables. For convenience, the response categories for item j are represented by consecutive integers, $1, 2, \ldots, R_j$. Thus, the data may be viewed as a J-way contingency table where the total number of cells is the product $R_1 \cdot R_2 \cdots R_J$. Assuming C latent classes, the mathematical model for LCA can be represented as:

$$\Pr(Y_i) = \sum_{c=1}^{C} \theta_c \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr}^{\delta_{ijr}} \tag{1}$$

The theoretical proportions of respondents in the latent classes are θ_c for $c = 1, \ldots, C$ with, of course, the sum of the proportions being one. Also, α_{cjr} is the theoretical conditional probability for response r to variable j given membership in latent class c. The terms, δ_{ijr}, are indicators that allow for the inclusion of appropriate conditional probabilities based on the responses, i.e.:

$$\delta_{ijr} = \begin{cases} 1 & \mathit{iff}\ y_{ij} = r \\ 0 & \mathit{otherwise} \end{cases} \tag{2}$$

Within any specific latent class, c, the probability for a response is:

$$\Pr(Y_i|c) = \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr}^{\delta ijr} \tag{3}$$

The product of conditional probabilities in equation (3) is based on the assumption that responses to the variables are independent *within latent classes*. To exemplify this, consider three dichotomous (1, 2) variables, two latent classes and the response {121} for some particular respondent. Within latent class 1, the theoretical probability for this response is the product $\alpha_{111}\alpha_{122}\alpha_{131}$ and within latent class 2 the probability is $\alpha_{211}\alpha_{222}\alpha_{231}$. Then, the unconditional probability for this response is the weighted sum, $\theta_1\alpha_{111}\alpha_{122}\alpha_{131} + (1-\theta_1)$ $\alpha_{211}\alpha_{222}\alpha_{231}$ where $\theta_2 = 1 - \theta_1$. Given the model in equation [1], the likelihood for a sample of n respondents is

$$\lambda = \prod_{i=1}^{n} \sum_{c=1}^{C} \theta_c \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr}^{\delta ijr} \tag{4}$$

and the logarithm of this likelihood is:

$$\Lambda = \sum_{i=1}^{n} \ln \left\{ \sum_{c=1}^{C} \theta_c \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr}^{\delta ijr} \right\} \tag{5}$$

Subject to identification conditions discussed below, maximum likelihood estimates for the parameters in equation (1) can be derived by computing partial derivatives of Λ with respect to the parameters, setting these partial derivatives to 0 and solving simultaneously. If the latent class proportion for the Cth class is rewritten as $\theta_C = 1 - \sum_{c=1}^{C-1} \theta_c$, the partial derivative with respect to the latent class proportion for class c is:

$$\frac{\delta\Lambda}{\delta\theta_c} = \sum_{i=1}^{n} \ln \Pr(Y_i|c) - \sum_{i=1}^{n} \ln \Pr(Y_i|C) = 0 \tag{6}$$

Similarly, a partial derivative with respect to a conditional probability for an item is of the form:

$$\frac{\delta\Lambda}{\delta\alpha_{cjr}} = \sum_{i=1}^{n} \ln \left[\theta_c \frac{\delta \Pr(Y_i|c)}{\delta\alpha_{cjr}} \right] = 0 \tag{7}$$

Although the partial derivatives in equations (6) and (7) are relatively easy to compute, they are nonlinear in the parameters and must be solved by iterative procedures. Various microcomputer programs are available for LCA, including LEM (Vermunt, 1997) that is based on an estimation-maximization (EM) algorithm, Latent Gold (Vermunt and Magidson, 2000) that incorporates Bayesian methods and MPlus (Muthén and Muthén, 1998) that uses Newton methods. The likelihood equation (4) is written on the assumption that the sample of respondents is a simple random sample from some population. However, for data arising from complex survey designs that incorporate clusters and sampling weights these methods must be modified as described by Patterson, Dayton and Graubard (2002). They define a pseudo-log-likelihood that incorporates sampling weights, w$_i$, for each respondent:

$$\Lambda_w = \sum_{i=1}^{n} w_i \ln \sum_{c=1}^{C} \theta_c \Pr(Y_i|c) \tag{8}$$

These sampling weights are intended to compensate for under- or oversampling strata, nonresponse and related factors (see Kish, 1965; Kish and Frankel, 1974; Kalton, 1989). As far as estimation *per se* is concerned, incorporating sampling weights into LCA is relatively straightforward and programs such as MPlus and Latent Gold have this capability. More difficult issues revolve around obtaining proper estimates for standard errors and setting up appropriate significance tests for data arising from complex surveys. As noted in Patterson, Dayton and Graubard (2002), this is still an active research area in LCA.

3 Model fit

Assessing the fit of a model to categorical data often entails a comparison between observed frequencies and expected frequencies where the latter are derived by substituting maximum-likelihood estimates for parameters in the theoretical model. This approach is practical unless the number of variables and/or numbers of categories for the variables becomes excessively large and data are sparse (i.e., many 0 and near-0 cell frequencies occur). There are three different chi-square goodness-of-fit statistics in common use: Pearson, likelihood-ratio, and Read-Cressie. All three statistics involve observed cell frequencies, F_t for $t = 1 \ldots, T$, and expected frequencies, $\hat{F}_t$ for $t = 1 \ldots, T$ where the number of cells is $T = \prod_{j=1}^{J} R_j$. Assuming a total sample size of n, the expected cell frequency for a cell with observation Y$_i$ is:

$$\hat{F}_t = n \cdot P(Y_i)_{i \in t} \tag{9}$$

The degrees of freedom for all three chi-square goodness-of-fit statistics are equal to: $T - p - 1$, where p is the number of independent parameters estimated when fitting the latent class model (e.g., for four dichotomous variables and two latent classes, the value of p is 9 and comprises one latent class proportion, four conditional probabilities for the variables in the first latent class, and four more conditional probabilities for the variables in the second latent class). As discussed below, these degrees of freedom are based on the assumption that the model is identified and unrestricted. For nonidentified models, only restricted solutions are possible and degrees of freedom must be adjusted accordingly.

The Pearson statistic, X^2, is based on the arithmetic difference between observed and expected frequencies:

$$X^2 = \sum_{t=1}^{T} \frac{(F_t - \hat{F}_t)^2}{\hat{F}_t} \tag{10}$$

The likelihood-ratio statistic, G^2 (sometimes denoted L^2), is based on the natural logarithm of ratios of observed and expected frequencies:

$$G^2 = 2\sum_{t=1}^{T} F_t \cdot \log_e \frac{F_t}{\hat{F}_t} \tag{11}$$

The Read-Cressie statistic, I^2, is a so-called power-divergence chi-square test (Read and Cressie, 1988) that is intended to be less sensitive to sparse data than either X^2 or G^2. It contains an adjustable constant, λ, that, when set to 0 yields G^2 and when set to 1 yields X^2. The value most frequently used for λ is 2/3 (e.g., this is the value used by the program LEM).

$$I^2 = \frac{2}{\lambda(1+\lambda)} \sum_{t=1}^{T} F_t \cdot \left[\left(\frac{F_t}{\hat{F}_t} \right)^{\lambda} - 1 \right] \tag{12}$$

In practice, X^2, G^2 and I^2 are often very similar in value but when X^2 and G^2 differ substantially it may be more appropriate to use I^2 as a measure of fit. However, it should be kept in mind that sparse data result in major distributional disturbances for all of these chi-square statistics since they are derived, in theory, from asymptotic properties associated with contingency tables. Also, it should be noted that accepting the null hypothesis for a goodness-of-fit test merely indicates that discrepancies between what is expected on the basis of a model and the data are within acceptable limits of chance and that many different models may provide "good" fit for a given dataset.

Given the limitations of goodness-of-fit tests and the fact that with large sample sizes the discrepancies between observed and expected frequencies can, subjectively, appear relatively small, various descriptive measures have been developed for use in LCA. A useful measure is the Index of Dissimilarity that ranges from 0 to 1 and is based on the absolute discrepancies between observed and expected frequencies:

$$I_D = \sum_{t=1}^{T} (|F_t - \hat{F}_t|)/(2n) \tag{13}$$

In practice, "satisfactory" fit is suggested by values of I_D less than .05. Also, the index of dissimilarity can be used to compare alternative models. Another interesting descriptive measure is the two-point mixture index of model fit, π^*, developed by Rudas, Clogg and Lindsay (1994). This index is the minimum proportion of respondents that must be deleted from the dataset in order to achieve perfect fit (i.e., $G^2 = 0$) for a given model. Although it is easily interpretable, there is no simple computational approach (see Dayton, 2003 for not-so-simple approaches) and is not currently available in latent class programs.

An issue of concern whenever applying advanced statistical models to real data is identification of the model. An identified model yields unique maximum likelihood estimates whereas a nonidentified model requires restrictions on the parameters in order to arrive at unique estimates. Trivially, one cannot estimate more independent parameters from a dataset than there are sufficient statistics for estimation. In the case of latent class analysis, the upper limit on the number of parameters that can be estimated is given by the number of potential data cells (e.g., 2^J for J dichotomous variables). Obvious identification issues are associated with the usual restrictions on probabilities; i.e., latent class proportions and conditional probabilities for variables sum to one as indicated above. However, these restrictions are very easy to incorporate into parameter estimation procedures. There is also the issue described as label switching which refers to the fact that, for example, interchanging what is labeled as the first versus the second latent class places the solution in different, albeit symmetric, locations in the parameter space (e.g., a split of .6, .4 versus .4, .6 for the latent class proportions). This, again, is not a serious identification issue but must be kept in mind when comparing solutions for two different groups of respondents such as males and females. Thus, the solutions may be similar but the first latent class estimated for males may be the equivalent to the second latent class estimated for females. More serious identification issues are associated with situations in which there is, in fact, no unique solution unless explicit restrictions are imposed on the estimates. A classic example involves three latent classes for four dichotomous variables. The number of cells is $2^4 = 16$ and it would seem that models with up to 15 parameters could be estimated with positive degrees of freedom remaining for assessing fit. However, an unrestricted three-class model based, apparently, on $2 + 3(4) = 14$ independent parameters is not identified and requires one restriction to yield unique estimates. This restriction, for example, could involve setting one the conditional probabilities for a variable to one. There is no straightforward method for ascertaining whether or not a given latent class model is identified. In theory, for an identified model, the asymptotic variance–covariance matrix for maximum-likelihood estimators is of full rank but, in general, this matrix can not be found by analytical methods and must be approximated numerically. Programs such as LEM compute such approximations but may fail to uncover an unidentified model. The best advice is to rerun the analysis with several different starting values for the iterative solution. If identical results are found, it is reasonable to assume that the model is identified unless there is some consistent boundary value (e.g., 1) for the conditional probabilities of one or more variables.

4 Model comparisons

Most applications of LCA entail fitting more than one model to a set of data. For nested models, with a very important exception as noted below, differences in chi-square goodness-of-fit tests (i.e., usually G^2 values) can be used to assess the statistical significance of the difference in fit with degrees of freedom equal to the difference in degrees of freedom for the models being compared. Thus, it can be decided whether or not the more complex model (i.e., the model with more independent parameters being estimated) provides better fit to the data than the simpler model. Note that the better fitting model may not actually represent "good" fit to the data but is simply the better of the models being compared.

The major exception to the use of difference chi-square tests is for the case of models with different numbers of latent classes which is, unfortunately, one of the cases of major interest in LCA. If one fits, say, two-class, three-class and four-class models to a set of data, these models are, in fact, nested. Despite this, the differences in X^2 or G^2 values are not distributed as theoretical chi-square variates. This is a general result that applies to comparing mixture models based on different numbers of components, including mixtures of normal distributions, mixtures of Poisson distributions, etc. (Titterington, Smith and Makov, 1985).

Although the reason for the failure of these tests is technically complex, it arises because, in moving from the more complex to the less complex model, the latent class proportion for the more complex model is constrained to zero, which is a boundary of the parameter space (Bishop, Fienberg and Holland, 1975).

An alternative for selecting a "best" model is based on notions from information theory. Akaike (1973; 1974) proposed an estimate for the Kullback and Leibler (1951) information measure that can be used for ordering alternate models that are estimated from the same data. Akaike interprets Kullback-Leibler information as a measure of the distance between the "true" model for data and models actually being estimated from the data. His estimate, AIC, may be viewed as a penalized form of the log-likelihood for a mode and is of the form $AIC = -2Ln(\lambda) + 2p = -2\Lambda + 2p$. Then, the decision-making strategy is to compute the AIC statistic for each model under consideration and select the model with min(AIC) as the preferred model *among those being compared*. An advantage of the min(AIC) strategy is that the models may be nested or non-nested. Various related measures have been proposed as alternatives to Akaike AIC. In general, these measures incorporate different (usually heavier) penalty terms and include Schwarz (1978) BIC with penalty term $\log_e(n)*p$ and Bozdogan (1987) CAIC with penalty term $[\log_e(n)+1]*p$. In this chapter, we utilize both the Akaike AIC and Schwarz BIC for model comparisons but suggest that readers consider the merits of alternatives (see Dayton, 1999).

5 Unconstrained LCA

One major application of LCA is to look for relatively homogeneous latent subgroups of respondents. As such, LCA may be viewed as a clustering algorithm that identifies "fuzzy" clusters. That is, as shown below, respondents may be classified into the latent classes on a *post hoc* basis but the assignment is probabilistic rather than deterministic. To illustrate this approach, we use data for 27,516 respondents to five survey items dealing with abortion taken from the General Social Survey (GSS) for the years 1972 through 1998. The items, in the order presented below, dealt with whether or not the respondent would favor allowing a woman to receive a legal abortion if: (1) there is a strong chance of serious defect in the baby; (2) she is married and does not want any more children; (3) the woman's own health is seriously endangered by the pregnancy; (4) she became pregnant as a result of rape; and (5) she is not married and does not want to marry the man. Frequencies for the 32 response patterns are displayed in Table 22.1 and latent class proportions, along with various summary statistics, based on fitting one to five latent classes are summarized in Table 22.2. As often true for large sample sizes, the various fit statistics present a complex picture that requires careful interpretation. First, only the five-class model results in a nonsignificant likelihood-ratio chi-square statistic and this model also satisfies the min(AIC) criterion. Second, the four-class model has min(BIC) and both the four- and five-class models have very small values for the Index of Discrepancy (.0022 and .0004, respectively). And, third, the latent class proportions for the three largest classes for the three- and four-class models are very similar. Given the small proportion in the final latent class for the four-class model, it seems reasonable to interpret the three-class model with the caveat that relatively small additional clusters of respondents might be reliable and of interest to some researchers. A plot of the three-class solution (Figure 22.1) shows a distinct pattern for each class. The largest class, about 49%, essentially agrees that an abortion for any of the five reasons should be legal. A relatively small class, about 13%, opposes an abortion for any of the given reasons but with opposition being somewhat abated for reasons of the mother's health

Table 22.1 Frequencies for five GSS abortion items

Defect	*No more*	*Health*	*Rape*	*Single*	*Observed frequency*
YES	YES	YES	YES	YES	11,212
YES	YES	YES	YES	NO	1265
YES	YES	YES	NO	YES	21
YES	YES	YES	NO	NO	79
YES	YES	NO	YES	YES	36
YES	YES	NO	YES	NO	20
YES	YES	NO	NO	YES	5
YES	YES	NO	NO	NO	4
YES	NO	YES	YES	YES	1471
YES	NO	YES	YES	NO	6859
YES	NO	YES	NO	YES	42
YES	NO	YES	NO	NO	1254
YES	NO	NO	YES	YES	17
YES	NO	NO	YES	NO	169
YES	NO	NO	NO	YES	3
YES	NO	NO	NO	NO	124
NO	YES	YES	YES	YES	68
NO	YES	YES	YES	NO	50
NO	YES	YES	NO	YES	7
NO	YES	YES	NO	NO	18
NO	YES	NO	YES	YES	21
NO	YES	NO	YES	NO	16
NO	YES	NO	NO	YES	2
NO	YES	NO	NO	NO	21
NO	NO	YES	YES	YES	95
NO	NO	YES	YES	NO	1206
NO	NO	YES	NO	YES	7
NO	NO	YES	NO	NO	1081
NO	NO	NO	YES	YES	14
NO	NO	NO	YES	NO	273
NO	NO	NO	NO	YES	4
NO	NO	NO	NO	NO	2052
					27,516

Table 22.2 Latent classes fitted to five GSS abortion items

# Latent classes	G^2	*DF*	*p-value*	*LC proportions*	I_D	*AIC*	*BIC*
1	43930.183	26	0.000	1.00	0.4779	145241.9	145283.0
2	8781.974	20	0.000	.517, .483	0.1619	110105.7	110196.2
3	256.684	14	0.000	.486, .387, .127	0.0060	101592.4	101732.2
4	25.582	8	0.001	.486, .371, .126, .017	0.0022	101373.3	**101562.4**
5	1.820	2	0.400	.484, .312, .127, .062, .014	0.0004	**101361.6**	101600.0

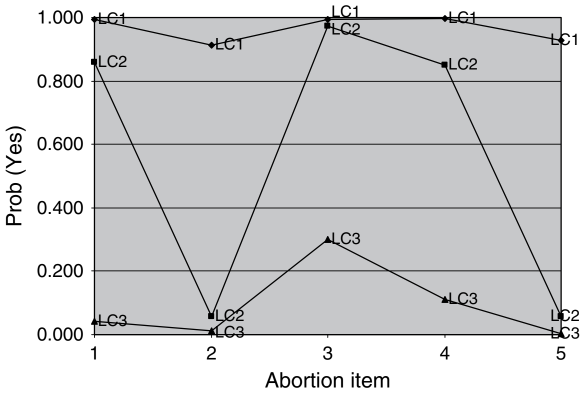


Figure 22.1 Three-class solution for GSS abortion items

or rape. Finally, a class representing about 37% of respondents has sharply divided opinions, being favorable toward reasons of a birth defect, mother's health or rape, but being unfavorable toward reasons of not wanting more children or being unmarried. It is interesting to note that, by combining the latent classes, one could characterize the sample as favorable to abortion (combine classes 1 and 2) or unfavorable to abortion (combine classes 2 and 3).

Note that the latent class proportions displayed in Figure 22.1 are consistent with the idea that the three latent classes are ordered. That is, all conditional probabilities for the first class are (equal to or) larger than those for the second class and all conditional probabilities for the second class are (equal to or) larger than those for the third class. This ordering occurred naturally given this dataset but order-restricted analyses can be imposed using options available in LCA programs such as LEM or Latent Gold.

Given maximum likelihood estimates for latent class proportions and conditional probabilities for the variables, Bayes theorem can be used to classify respondents into the latent classes. In the context of LCA, the theorem takes the form

$$\Pr(c|\mathbf{Y}_i) \propto \theta_c \cdot \Pr(\mathbf{Y}_i|c) \tag{14}$$

which means that for a given response, $\mathrm{Y_i}$, the (posterior) probability for latent class c *is proportional to* the latent class proportion times the likelihood associated with that response. Then, classification is carried by assigning a response to the latent class for which the posterior probability is largest (i.e., the model posterior class). Table 22.3 displays the classifications for the three class solution based on the five GSS abortion items. If the frequencies for the modal classes (in bold in the table) are summed and converted to proportions, they are .510, .360

Table 22.3 Bayes classifications for five GSS abortion items

						Bayes probabilities		
Defect	*No more*	*Health*	*Rape*	*Single*	*Observed frequency*	*Class 1*	*Class 2*	*Class 3*
YES	YES	YES	YES	YES	11,212	**0.998**	0.002	0.000
YES	YES	YES	YES	NO	1265	**0.683**	0.317	0.000
YES	YES	YES	NO	YES	21	**0.827**	0.173	0.000
YES	YES	YES	NO	NO	79	0.021	**0.971**	0.008
YES	YES	NO	YES	YES	36	**0.989**	0.011	0.000
YES	YES	NO	YES	NO	20	0.285	**0.704**	0.010
YES	YES	NO	NO	YES	5	0.461	**0.513**	0.026
YES	YES	NO	NO	NO	4	0.002	**0.596**	0.402

Table 22.3 (Continued)

						Bayes probabilities		
Defect	*No more*	*Health*	*Rape*	*Single*	*Observed frequency*	*Class 1*	*Class 2*	*Class 3*
YES	NO	YES	YES	YES	1471	**0.729**	0.271	0.000
YES	NO	YES	YES	NO	6859	0.012	**0.987**	0.001
YES	NO	YES	NO	YES	42	0.027	**0.971**	0.002
YES	NO	YES	NO	NO	1254	0.000	**0.967**	0.033
YES	NO	NO	YES	YES	17	0.335	**0.662**	0.003
YES	NO	NO	YES	NO	169	0.002	**0.937**	0.061
YES	NO	NO	NO	YES	3	0.004	**0.817**	0.179
YES	NO	NO	NO	NO	124	0.000	0.254	**0.746**
NO	YES	YES	YES	YES	68	**0.955**	0.045	0.000
NO	YES	YES	YES	NO	50	0.087	**0.893**	0.020
NO	YES	YES	NO	YES	7	0.167	**0.774**	0.060
NO	YES	YES	NO	NO	18	0.001	0.490	**0.510**
NO	YES	NO	YES	YES	21	**0.777**	0.194	0.029
NO	YES	NO	YES	NO	16	0.006	0.332	**0.662**
NO	YES	NO	NO	YES	2	0.005	0.128	**0.867**
NO	YES	NO	NO	NO	21	0.000	0.011	**0.989**
NO	NO	YES	YES	YES	95	0.108	**0.886**	0.007
NO	NO	YES	YES	NO	1206	0.001	**0.909**	0.091
NO	NO	YES	NO	YES	7	0.001	**0.747**	0.252
NO	NO	YES	NO	NO	1081	0.000	0.181	**0.819**
NO	NO	NO	YES	YES	14	0.014	**0.598**	0.388
NO	NO	NO	YES	NO	273	0.000	0.103	**0.897**
NO	NO	NO	NO	YES	4	0.000	0.033	**0.967**
NO	NO	NO	NO	NO	2052	0.000	0.003	**0.998**
					27,516			

and .131 which closely correspond to the latent class proportions (i.e., .486, .387 and .127). There is a relationship between classification into the latent classes and the count of Yes responses to the GSS abortion items but there are notable exceptions. In particular, counts of 4 and 5 are unique to latent class 1 and counts of 0 and 1 are unique to latent class 3, but counts of 3 occur in both latent classes 1 and 2 and counts of 2 occur in both latent classes 2 and 3. Thus, in this regard, the latent class analysis provides a more highly nuanced interpretation of the responses.

6 Multiple groups LCA

Comparisons among subgroups within a sample are often of interest. Typical comparison groups are males/females, age groups, ethnic groups, etc. The model for LCA in equation (1) can be adapted to accommodate grouping variables (Clogg and Goodman, 1984; 1985; Dayton and Macready, 2002). We consider only a single grouping variable but the method is easily extended to more complex stratifications. Letting an observation for a respondent in group g be $\mathbf{Y}_{ig} = \{y_{igj}\}$, the model becomes:

$$\Pr\left(Y_{ig}\right) = \sum_{c=1}^{C} \theta_{c|g} \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr|g}^{\delta_{igjr}} \tag{15}$$

Note that the latent class proportions, $\theta_{c|g}$, the item conditional probabilities, $\alpha_{cjr|g}$, as well as the indicators, δ_{igjr} include a subscript for group membership. The grouping latent class model in equation (15) is referred to as a heterogeneous model. In the heterogeneous model, the latent classes may or may not have a consistent interpretation across groups. In typical applications, this heterogeneous model is compared to a partially homogeneous model as well as to a completely homogeneous model. For a model with G groups, the partially homogeneous model is defined by the restrictions

$$\alpha_{cjr|g}^{\delta_{igjr}} = \alpha_{cjr}^{\delta_{isjr}} \; for\; g = 1, \ldots , G \tag{16}$$

For this model, the sizes of the latent classes are allowed to vary across groups but the conditional probabilities for the variables that characterize the nature of the latent classes remain the same so that the interpretation of the classes can not vary from group to group. On the other hand, the homogeneous model is defined by both the restrictions in equation (16) and

$$\theta_{c|g} = \theta_c \; for\; g = 1, \ldots , G \tag{17}$$

In effect, grouping is ignored when these restrictions are applied. Statistical comparisons among heterogeneous, partially homogeneous, and completely homogeneous models can be carried out using difference chi-square tests since these models are nested and do not involve setting latent class proportion to boundary values.

To illustrate these methods, the dataset for the five GSS abortion items was divided into younger (age 42 and below) versus older (age 43 and above) age groups. This division reduces the sample size to 27,697 because of missing data for age and represents about an even split of the respondents (53% versus 47%). On an item-by-item basis, the younger group uniformly tends to give more yes responses to the abortion items. Homogeneous, partial homogeneous, and heterogeneous models are summarized in Table 22.4. Note that the latent class proportions are presented in the same order as for the earlier analyses; i.e., the first class is relatively favorable to abortions for the stated reasons, the third class is relatively opposed and the second class is opposed for what might be characterized as nonmedical reasons. Although all of the chi-square fit statistics suggest lack of model fit, the indices of discrepancy are about 2% or less for both the partial homogeneous and heterogeneous models. If we choose to interpret the partial homogeneous model, this is supported by the fact that this model satisfies the min(BIC) criterion. As shown in Table 22.4, the younger group has a higher proportion in the favorable class, a lower proportion in the unfavorable class, and about the same proportion in the second class when compared with the older group.

Table 22.4 Grouping latent class models fitted to five GSS abortion items grouped by age LC proportions

Model	*G^2*	*DF*	*p-value*	*Younger group*	*Older group*	*I_D*	*AIC*	*BIC*
Homogeneous	539.872	45	0.000	.491, .384, .126	.491, .384, .126	0.0426	140067.8	140215.9
Partial homogeneous	373.050	43	0.000	.522, .374, .104	.455, .396, .149	0.0209	139902.9	**140059.3**
Heterogeneous	274.179	28	0.000	.518, .375, .107	.461, .392, .147	0.0063	**139834.1**	140113.8

7 Scaling models

One of the first applications of restricted latent class models involved assessing the fit of linear (Guttman) scales to attitude rating items. These same models have applications to learning hierarchies and to other outcomes that might be expected to occur in some specific sequence over time periods (e.g., developmental tasks). Proctor (1970), Goodman (1974; 1975) and Dayton and Macready (1976), among others, presented models in which each scale type was represented by a separate latent class and the conditional probabilities for the variables (e.g., items) were suitably restricted to define a hierarchic or sequential structure. This chapter provides some basic concepts on this topic; more complete coverage can be found in Dayton (1999).

Consider, for example, three attitude items representing increasing degrees of positive opinion on some social topic. If responses (Y = yes, N = no) to these items conformed to a linear scale, the permissible response patterns would be NNN, YNN, YYN, YYY, whereas a response like NYN would be inconsistent with a linear scale. In practice, of course, such inconsistent responses do occur and the issue is whether or not the linear scale is a reasonable approximation for observed data. Latent class scaling models assume that each permissible response pattern is associated with a latent class, so for three items, as above, there would be four latent classes. In order to identify a latent class with a specific permissible response pattern, restrictions are imposed on the conditional probabilities for the items. A very simple model proposed by Proctor (1970) assumes response errors at a constant rate across items. Thus, respondents in the latent class corresponding to the pattern NNN may, in fact, show any of the other seven possible response patterns if one, two, or three response errors are made. Of course, a respondent would give the NNN response if no error were made. In this way, respondents in any of the four latent classes representing the four permissible response patterns may give any possible response with greater or lesser probability. The formal restrictions on the item conditional probabilities are:

$$\alpha_e = \alpha_{111} = \alpha_{121} = \alpha_{131} = \alpha_{221} = \alpha_{231} = \alpha_{331}$$
$$= \alpha_{212} = \alpha_{312} = \alpha_{322} = \alpha_{412} = \alpha_{422} = \alpha_{432}$$

Note that α_e can be viewed as a constant error rate that, depending on the permissible response pattern, corresponds to a response of Y occurring when the permissible response is N or a response of N occurring when the permissible response is Y. Thus, for example, if a respondent comes from the latent class corresponding to YNN, the response NNN has probability $\alpha_e(1-\alpha_e)^2$ since there is one response error and two non-error responses. Similarly, the response YNN has probability $(1-\alpha_e)^3$ representing three non-errors and the response NYY has probability α_e^3 representing three response errors.

The Proctor model was extended by Dayton and Macready (1976) to include different types of response errors. The intrusion-omission error model posits two types of errors corresponding to Y replacing N (intrusion) and N replacing Y (omission). The formal restrictions on the item conditional probabilities are:

$$\alpha_I = \alpha_{111} = \alpha_{121} = \alpha_{131} = \alpha_{221} = \alpha_{231} = \alpha_{331} \text{ and}$$
$$\alpha_O = \alpha_{212} = \alpha_{312} = \alpha_{322} = \alpha_{412} = \alpha_{422} = \alpha_{432}$$

Other error models include item specific errors:

$$\alpha_{e1} = \alpha_{111} = \alpha_{212} = \alpha_{312} = \alpha_{412};$$
$$\alpha_{e2} = \alpha_{121} = \alpha_{221} = \alpha_{322} = \alpha_{422}; \text{ and}$$
$$\alpha_{e3} = \alpha_{131} = \alpha_{231} = \alpha_{331} = \alpha_{432}$$

as well as latent-class specific errors:

$$\alpha_{c1} = \alpha_{111} = \alpha_{121} = \alpha_{131}; \alpha_{c2} = \alpha_{221} =$$
$$\alpha_{231} = \alpha_{212}; \alpha_{c3} = \alpha_{331} = \alpha_{312} = \alpha_{322}; \text{ and}$$
$$\alpha_{c4} = \alpha_{412} = \alpha_{422} = \alpha_{432}$$

Although the GSS abortion items were neither designed nor selected to be consistent with a linear scale, we explore some scaling models to how they compare in fit with the unrestricted models fitted earlier. Based on purely subjective judgments gleaned from inspecting Figure 22.1, the item ordering 3, 4, 1, 5, 2 was selected. Thus, the permissible response patterns are: NNNNN, NNYNN, NNYYN, YNYYN, YNYYY and YYYYY. Table 22.5 summarizes results from fitting four error models. The best fitting, among these generally poor fitting models, is the item-specific error model that meets both the min(AIC) and min(BIC) criteria, as well as having the most acceptable I_D value. Note, however, that these values, as well as the G^2 statistic, are not nearly as acceptable as those found for an unrestricted three-class model (Table 22.1). The estimated latent class proportions for the permissible response patterns in the order above are: .090, .040, .046, .339, .021 and .464. The first and fifth classes correspond to the extreme groups for the unrestricted three-class model and show similar latent class proportions to that solution. Note that the estimated error rates for two of the five classes for the latent-class specific error model went to boundary values of 0, which suggests identification issues for this model.

8 Covariate LCA

Although grouping variables provides a useful approach to incorporating additional manifest variables into LCA, it has definite limitations. First, if an outside variable is continuous, rather than categorical (e.g., age of respondent), then it is necessary to create groups using some more-or-less arbitrary cut-values as was illustrated above. Second, only a relatively few grouping variables can be accommodated in an analysis. If large numbers of grouping variables are used, the cell frequencies become small, resulting in unstable subgroup analyses. And, third, LCA with grouping variables can be extremely complex in terms of the number of parameters that are estimated. Dayton and Macready (1988) proposed covariate LCA in which a logistic regression model is written for the relationship between latent class membership and one or more covariates (actually, they proposed a more general model, but the logistic model is most widely used in practice). The covariates can be continuous, categorical using recoding when necessary (e.g., dummy-variable coding), and/or products of covariates in order to model interactions. In effect, all of the modeling possibilities available in ordinary multiple regression and logistic multiple regression become available in the context of LCA.

Table 22.5 Linear scale fitted to five GSS abortion items

Error model	*G^2*	*DF*	*p-value*	*Error rates*	*I_D*	*AIC*	*BIC*
Proctor	4202.110	25	0.000	0.040	0.095	105515.9	105565.2
Intrusion-omission	3869.276	24	0.000	.082, .024	0.087	105185.0	105242.6
Item specific	3073.560	21	0.000	.015, .039, .013, .067, .062	0.078	**104395.3**	**104477.5**
Latent class specific	6980.709	20	0.000	.004, .195, .000*, .077, .000*, .004	0.099	108304.5	108394.9

* Conditional probability at boundary value; identification issues

Although covariate LCA can be generalized as shown below, we begin by assuming that there are just two latent classes and a single covariate, Z. The model for the regression of latent class membership on the covariate is:

$$\theta_{1|Z_i} = g(Z_i|\beta) = \frac{e^{\beta_0+\beta_1 Z_i}}{1+e^{\beta_0+\beta_1 Z_i}} \tag{18}$$

In log-odds form, the model can be written:

$$\ln\left(\frac{\theta_{1|Z_i}}{1-\theta_{1|Z_i}}\right) = \beta_0+\beta_1 Z_i \tag{19}$$

Note that this is exactly the same model posited in ordinary logistic regression analysis except that membership in the latent classes is unknown rather than manifest. Combining the model in equation (18) with the latent class model for two latent classes yields:

$$\Pr(Y_i|Z_i) = \sum_{c=1}^{2} \theta_{c|Z_i} \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{cjr}^{\delta_{ijr}} \tag{20}$$

Note that the probability of latent class membership is dependent on the covariate but that conditional probabilities for the variables are independent of the variables. In the terminology of models with grouping variables, this is a partially homogeneous model. That is, the latent structure defined by the conditional probabilities for the variables is assumed to be constant over the different values for the covariate.

Covariate LCA can accommodate multiple covariates and cases with three or more latent classes. For multiple covariates, the obvious modification is to expand the additive model for log-odds:

$$\ln\left(\frac{\theta_{1|Z_i}}{1-\theta_{1|Z_i}}\right) = \beta_0+\beta_1 Z_{1i}+\cdots+\beta_p Z_{pi} \tag{21}$$

where Z_i is a vector of p covariates. One approach to extending the model to more than two latent classes is to select one of the classes as a reference (usually the last class, C) and then create log-odds models comparing each of the remaining classes with the reference class. Using this coding, by default, the logistic regression coefficients for the reference class are each equal to 0. Then there are C-1 log-odds models analogous to equation (21) where the logistic regression models for classes c = 1…,C-1 are of the form:

$$\theta_{c|Z_i} = g(Z_i|\beta) = \frac{e^{\beta_{0c}+\beta_{1c} Z_i}}{1+\sum_{c=1}^{C-1} e^{\beta_{0c}+\beta_{1c} Z_i}} \tag{22}$$

and for class C is:

$$\theta_{C|Z_i} = g(Z_i|\beta) = \frac{1}{1+\sum_{c=1}^{C-1} e^{\beta_{0c}+\beta_{1c} Z_i}} \tag{23}$$

For illustration, we return to the five GSS abortion items where a three-class model provided reasonable fit. Using age as a continuous covariate, the estimated conditional probabilities for the five items were very similar to those shown in Figure 22.1 and are not summarized here. Thus, the structure of reported abortion attitudes is essentially the same as found from the LCA without age as a covariate. The logistic regression coefficients for the first two latent class proportions were estimated as $\beta_{01} = 2.084, \beta_{11} = -.016$ and $\beta_{02} = 1.639, \beta_{12} = -.011$. Figure 22.2 displays the distinctive patterns for the three latent classes. With increasing age, the expected percentage in the first latent class that is relatively favorable to abortion declines steadily from about 55% to about 40%. On the other hand, the third latent class that is relatively unfavorable to abortion increases from less than 10% to more than 20% with increasing age. The third latent class that tends to oppose abortion for nonmedical reasons is relatively stable over the range of ages in the GSS database.

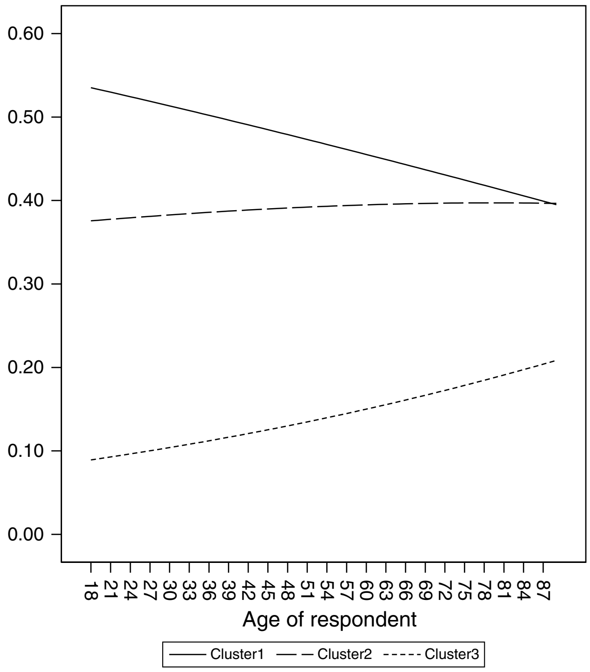


Figure 22.2 GSS abortion item cluster profiles with age as covariate

9 Software notes

All analyses reported in this paper were run with Latent Gold (Vermunt and Magidson, 2000) from datasets in SPSS file. All Bayes constants were set to 0 which results in maximum likelihood estimates rather than posterior-mode Bayes estimates. Equivalent analyses could have been generated using LEM (Vermunt, 1997) but this could have required the creation of new datasets since inputting SPSS datasets is not an option with LEM. The GSS abortion items were taken from public-access databases maintained by the National Opinion Research Center in Chicago, Illinois, at the website http://webapp.icpsr.umich.edu/GSS/.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csake (eds), *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* AC-19: 716–723.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. London: Charles Griffin & Co.

Bishop, Y., Fienberg, S. and Holland, P. (1975) *Discrete Multivariate Analysis*. MIT Press.

Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika,* 52: 345–370.

Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional tables. *Journal of the American Statistical Association*, 79: 762–771.

Clogg, C. C. and Goodman, L. A. (1985). Simultaneous latent structure analysis in several groups. In N. B. Tuma (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.

Dayton, C. M. (1999). *Latent Class Scaling Analysis*. Quantitative Applications in the Social Sciences Series No. 126. Thousand Oaks, CA: Sage.

Dayton, C. M. (2003). Applications and computational strategies for the two-point mixture index of fit. *British Journal of Mathematical and Statistical Psychology*, 56: 1–13.

Dayton, C. M. and Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. *Psychometrika,* 41: 189–204.

Dayton, C. M. and Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika,* 45: 343–356.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent class models. *Journal of the American Statistical Association*, 83: 173–178.

Dayton, C. M. and Macready, G. B. (2002). Use of categorical and continuous covariates in latent class analysis. In Allan McCutcheon and Jacques Hagenaars (eds), *Advances in Latent Class Modeling*. Cambridge, UK: Cambridge University Press.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, 61: 215–231.

Goodman, L. A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. *Journal of the American Statistical Association,* 70: 755–768.

Haberman, S. J. (1979). *Analysis of Quantitative Data*, Vol. 2. New York: Academic Press.

Hagenaars, J. A. (1990). *Categorical Longitudinal Data*. Newbury Park: Sage.

Hagenaars, J. A. and McCutcheon, A. L. (eds) (2002). *Applied Latent Class Analysis.* Cambridge, UK: Cambridge University Press.

Heinen, T. (1996). *Latent Class and Discrete Trait Models.* Advanced Quantitative Techniques in the Social Sciences Series 6. Thousand Oaks: Sage.

Kalton, G. (1989). Modeling considerations: Discussion from a survey sampling perspective. In D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh (eds), *Panel Survey*, pp. 575–585. New York: Wiley.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Kish, L. and Frankel, M. P. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, *Series B,* 36: 1–37.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics,* 22: 79–86.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Muthén, L. K. and Muthén, B. O. (1998). *Mplus: The Comprehensive Modeling Program for Applied Researchers, User's Guide*. Los Angeles, CA: Muthén & Muthén.

Patterson, B., Dayton, C. M. and Graubard, B. (2002). Latent class analysis of complex survey data: Application to dietary data. *Journal of the American Statistical Association*, 97: 721–729.

Proctor, C. H. (1970). A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika,* 35: 73–78.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Rindskopf, R. and Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine,* 5: 21–27.

Rost, J. and Langeheine, R. (eds) (1997). *Applications of Latent Trait and Latent Class Models in the Social Sciences*. New York: Waxmann.

Rudas, T., Clogg, C. C. and Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables*. Journal of the Royal Statistical Society, Series B,* 56: 623–639.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6: 461–464.

Spearman, C. E. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 5: 201–293.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Models*. New York: Wiley.

Vermunt, J. K. (1997), *The LEM user manual*. WORC Paper. Tilburg University, The Netherlands.

Vermunt, J. K. and Magidson, J. (2000). Latent class cluster analysis. In J. A. Hagenaars and A. L. McCutcheon (eds), *Advances in Latent Class Models*. Cambridge, UK: Cambridge University Press.

Von Eye, A. and Clogg, C. C. (eds) (1994). *Latent Variables Analysis: Applications for Developmental Research*. Thousand Oaks: Sage.

This page intentionally left blank

Chapter 23

Latent class models in longitudinal research

Jeroen K. Vermunt, Bac Tran and Jay Magidson

1 Introduction

This article presents a general framework for the analysis of discrete-time longitudinal data using latent class models. The encompassing model is the mixture latent Markov model, a latent class model with time-constant and time-varying discrete latent variables. The time-constant latent variables are used to deal with unobserved heterogeneity in the change process, whereas the time-varying discrete latent variables are used to correct for measurement error in the observed responses. By allowing for direct relationships between the latent states at consecutive time points, one obtains the typical Markovian transition or first-order autoregressive correlation structure. Moreover, each of three distinct submodels can include covariates, thus addressing separate important issues in longitudinal data analysis: observed and unobserved individual differences, autocorrelation, and spurious observed change resulting from measurement error.

It is shown that most of the existing latent class models for longitudinal data are restricted special cases of the mixture latent Markov model presented, which itself is an expanded version with covariates of the mixed Markov latent class model by Van de Pol and Langeheine (1990). The most relevant restricted special cases are mover-stayer models (Goodman, 1961), mixture Markov models (Poulsen, 1982), latent (or hidden) Markov models (Baum et al., 1970; Collins and Wugalter, 1992; Van de Pol and De Leeuw, 1986; Vermunt, Langeheine and Böckenholt, 1999; Wiggins, 1973), mixture growth models (Nagin, 1999; Muthén, 2004; Vermunt, 2006), and mixture latent growth models (Vermunt, 2003; 2006) for repeated measures, as well as the standard multiple-group latent class model for analyzing data from repeated cross-sections (Hagenaars, 1990).

The next section presents the mixture latent Markov model. Then we discuss its most important special cases and illustrate these with an empirical example. We end with a short discussion of various possible extensions of our approach. The first appendix provides details on parameter estimation using the Baum-Welch algorithm. The second appendix contains model setups for the syntax version of the Latent GOLD program (Vermunt and Magidson, 2005) that was used for estimating the example models.

2 The mixture latent Markov model

Assume that we have a longitudinal data set containing measurements for N subjects at $T+1$ occasions. The mixture latent Markov model is a model containing five types of variables: response variables, time-constant explanatory variables, time-varying explanatory variables, time-constant discrete latent variables, and time-varying discrete latent variables. For simplicity of exposition, we will assume that response variables are categorical, and that there is at most one time-constant and one time-varying latent variable. These are, however, not limitations of the framework we present which can be used with continuous response variables, multiple time-constant latent variables, and multiple time-varying latent variables. Our mixture latent Markov model is an expanded version of the mixed Markov latent class model proposed by Van de Pol and Langeheine (1990): it contains time-constant and time-varying covariates and it can be used when the number of time points is large.

Let y_{itj} denote the response of subject i at occasion t on response variable j, where $1 \leq i \leq N$, $0 \leq t \leq T$, $1 \leq j \leq J$, and $1 \leq y_{itk} \leq M_j$. Note that J is the total number of response variables and M_j the number of categories for response variable j. The vector of responses for subject i at occasion t is denoted as $\mathbf{y}_{it}$ and the vector of responses at all occasions as $\mathbf{y}_i$. The vector of time-constant and time-varying predictors at occasion t is denoted by $\mathbf{z}_i$ and $\mathbf{z}_{it}$, respectively. The time-constant and time-varying discrete latent variables are denoted by w and x_t, where $1 \leq w \leq L$ and $1 \leq x_t \leq K$. The latter implies that the number of categories of the two types of latent variables equal L and K, respectively. To make the distinction between the two types of latent variables clear, we will refer to w as a latent class and to x_t as a latent state.

The general model that we use as the starting point is the following mixture latent Markov model:

$$P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L}\sum_{x_0=1}^{K}\sum_{x_1=1}^{K}\ldots\sum_{x_T=1}^{K} P(w, x_0, x_1, \ldots, x_T|\mathbf{z}_i) \times P(\mathbf{y}_i|w, x_0, x_1, \ldots, x_T, \mathbf{z}_i) \quad (1)$$

with

$$P(w, x_0, x_1, \ldots, x_T|\mathbf{z}_i) = P(w|\mathbf{z}_i) \times P(x_0|w, \mathbf{z}_{i0}) \prod_{t=1}^{T} P(x_t|x_{t-1}, w, \mathbf{z}_{it}) \quad (2)$$

$$P(\mathbf{y}_i|w, x_0, x_1, \ldots, x_T, \mathbf{z}_i) = \prod_{t=0}^{T} P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it}) = \prod_{t=0}^{T}\prod_{j=1}^{J} P(y_{itj}|x_t, w, \mathbf{z}_{it}) \quad (3)$$

As many statistical models, the model in equation (1) describes $P(\mathbf{y}_i|\mathbf{z}_i)$, the (probability) density associated with responses of subject i conditional on his/her observed covariate values. The right-hand side of this equation shows that we are dealing with a mixture model containing 1 time constant latent variable and $T+1$ time-varying latent variables. The total number of mixture components (or latent classes) equals $L \cdot K^{T+1}$, which is the product of the number of categories of w and x_t for $t = 0, 1, 2, \ldots, T$. As in any mixture model, $P(\mathbf{y}_i|\mathbf{z}_i)$ is obtained as a weighted average of class-specific probability densities – here $P(\mathbf{y}_i|w, x_0, x_1, \ldots, x_T, \mathbf{z}_i)$ – where the (prior) class membership probabilities or mixture proportions – here $P(w, x_0, x_1, \ldots, x_T|\mathbf{z}_i)$ – serve as weights (Everitt and Hand, 1981; McLachlan and Peel, 2000).

Equations (2) and (3) show the specific structure assumed for the mixture proportion $P(w, x_0, x_1, \ldots, x_T|\mathbf{z}_i)$ and the class-specific densities $P(\mathbf{y}_i|w, x_0, x_1, \ldots, x_T, \mathbf{z}_i)$. The equation for $P(w, x_0, x_1, \ldots, x_T|\mathbf{z}_i)$ assumes that conditional on w and $\mathbf{z}_i$, x_t is associated only with x_{t-1} and x_{t+1} and thus not with the states occupied at the other time points – the well-known first-order Markov assumption. The equation for $P(\mathbf{y}_i|w, x_0, x_1, \ldots, x_t, \mathbf{z}_i)$ makes two assumptions: (1) conditionally on w, x_t, and $\mathbf{z}_{it}$, the J responses at occasion t are independent of the latent states and the responses at other time points, and (2) conditionally on w, x_t, and $\mathbf{z}_{it}$, the J responses at occasion time point t are mutually independent, which is referred to as the local independence assumption in latent class analysis (Goodman, 1974).

As can be seen from equations (2) and (3), the models of interest contain four different kinds of model probabilities:

- $P(w|\mathbf{z}_i)$ is the probability of belonging to a particular latent class conditional on a person's covariate values.
- $P(x_0|w, \mathbf{z}_{i0})$ is an initial-state probability; i.e., the probability of having a particular latent initial state conditional on an individual's class membership and covariate values at $t = 0$.
- $P(x_t|x_{t-1}, w, \mathbf{z}_{it})$ is a latent transition probability; i.e., the probability of being in a particular latent state at time point t conditional on the latent state at time point $t-1$, class membership, and time-varying covariate values.
- $P(y_{itj}|x_t, w, \mathbf{z}_{it})$ is a response probability, which is the probability of having a particular observed value on response variable j at time point t conditional on the latent state occupied at time point t, class membership w, and time-varying covariate values.

Typically, these four sets of probabilities will be parameterized and restricted by means of (logistic) regression models. This is especially useful when a model contains covariates, where time itself may be one of the time-varying covariates of main interest. In the empirical application presented below we will use such regression models. For extended discussions on logistic regression analysis, we refer to introductory texts on this topic (see, for example, Agresti, 2002; Menard, 2002; Vermunt, 1997).

The three key elements of the mixture latent Markov model described in equations (1), (2), and (3) are that it can take into account (1) unobserved heterogeneity, (2) autocorrelation, and (3) measurement error. Unobserved heterogeneity is captured by the time-constant latent variable w, autocorrelations are captured by the first-order Markov transition process in which the state at time point t may depend on the state at time point $t-1$, and measurement error or misclassification is accounted for allowing an imperfect relationship between the time-specific latent states x_t and the observed responses y_{itj}. Note that these are three of the main elements that should be taken into account in the analysis of longitudinal data; i.e., the interindividual variability in patterns of change, the tendency to stay in the same state between consecutive occasions, and spurious change resulting from measurement error in observed responses.

Parameters of the mixture latent Markov model can be estimated by means of maximum likelihood (ML). For that purpose, it is advisable to use a special variant of the expectation maximization (EM) algorithm that is usually referred to as the forward-backward or Baum-Welch algorithm (Baum et al., 1970; McDonald and Zucchini, 1997) which is described in detail in the first appendix. This special algorithm is needed because our model contains a potentially huge number of entries in the joint posterior latent distribution $P(w, x_0, x_1, \ldots, x_T|\mathbf{y}_i, \mathbf{z}_i)$, except in cases where T, L and K are all small. For example, in a fairly moderate sized situation where $T = 10$, $L = 2$ and $K = 3$, the number of entries in the joint posterior distribution already equals $2 \cdot 3^{11} = 354294$, a number which is impossible to process and store for all N subjects as has to be done within standard EM. The Baum-Welch algorithm circumvents the computating of this joint posterior distribution making use of the conditional independencies implied by the model. Vermunt (2003) proposed a slightly simplified version of the Baum-Welch algorithm for dealing with the multilevel latent class model, which when used for longitudinal data analysis is one of the special cases of the mixture latent Markov model described in the next section.

A common phenomenon in the analysis of longitudinal data is the occurrence of missing data. Subjects may have missing values either because they refused to participate at some occasions or because it is elected by the study design. A nice feature of the approach described here is that it can easily accommodate missing data in the ML estimation of the unknown model parameter. Let δ_{it} be an indicator variable taking on the value 1 if subject i provides information for occasion t and 0 if this information is missing. The only required change with missing data is the following modification of equation (3):

$$P(\mathbf{y}_i|w, x_0, x_1, \ldots, x_T, \mathbf{z}_i) = \prod_{t=0}^{T} [P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})]^{\delta_{it}}.$$

For $\delta_{it} = 1$, nothing changes compared to what we had before. However, for $\delta_{it} = 0$, the time-specific conditional density becomes 1, which means that the responses of a time point with missing values are skipped. Actually, for each pattern of missing data, we have a mixture latent Markov for a different set of occasions. Two limitations of the ML estimation procedure with missing values should be mentioned: (1) it can deal with missing values on response variables, but not with missing values on covariates, and (2) it assumes that the missing data are missing at random (MAR). The first limitation may be problematic when there are time-varying covariates for which the values are also missing. However, in various special cases discussed below – the ones that do not use a transition structure – it is not a problem if time-varying covariates are missing for the time points in which the responses are missing. The second limitation concerns the assumed missing data mechanism: MAR is the least restrictive mechanism under which ML estimation can be used without the need of specifying the exact mechanism causing the missing data; i.e., under which the missing data mechanism is ignorable for likelihood-based inference (Little and Rubin, 1987; Schafer, 1997). It is possible to relax the MAR assumption by explicitly defining a not-missing-at-random (NMAR) mechanism as a part of the model to be estimated (Fay, 1986; Vermunt, 1997).

An issue strongly related to missing data is the one of unequally spaced measurement occasions. As long as the model parameters defining the transition probability are assumed to be occasion specific, no special arrangements are needed. If this is not the case, unequally spaced measurements can be handled by defining a grid of equally spaced time points containing all measurement occasions. Using this technique, the information on the extraneous occasions can be treated as missing data for all subjects. An alternative is to use a continuous-time rather than a discrete time framework (Böckenholt, 2005), which can be seen as the limiting case in which the elapsed time between consecutive time points in the grid approaches zero.

Another issue related to missing data is the choice of the time variable and the corresponding starting point of the process. The most common approach is to use calender time as the time variable and the first measurement occasion as $t = 0$, but one may, for example, also use age as the relevant time variable, as we do in the empirical example. Although children's ages at the first measurement vary between 11 and 17, we use age 11 as $t = 0$. This implies that for a child that is 12 years of age information at $t = 0$ is treated as missing, for a child that is 13 years of age information a $t = 0$ and $t = 1$ is treated as missing, etc.

3 The most important special cases

Table 23.1 lists the various special cases that can be derived from the mixture latent Markov model defined in equations (1)–(3) by assuming that one or more of its three elements –

Table 23.1 Classification of latent class models for longitudinal research

	Model name	*Transition structure*	*Unobserved heterogeneity*	*Measurement error*
I	Mixture latent Markov	yes	yes	yes
II	Mixture Markov	yes	yes	no
III	Latent Markov	yes	no	yes
IV	Standard Markov*	yes	no	no
V	Mixture latent growth	no	yes	yes
VI	Mixture growth	no	yes	no
VII	Standard latent class	no	no	yes
VIII	Independence*	no	no	no

*This model is not a latent class model.

transition structure, measurement error, and unobserved heterogeneity – is not present or needs to be ignored because the data is not informative enough to deal with it. Models I–III and V–VII are latent class models, but IV and VIII are not. Model VII differs from models I–VI in that it is a model for repeated cross-sectional data rather than a model for panel data. Below we describe the various special cases in more detail.

3.1 Mixture latent Markov

First of all, it is possible to define simpler versions of the mixture latent Markov model itself. Actually, the mixed Markov latent class model proposed by Van de Pol and Langeheine (1990) which served as an inspiration for our model is the special case of our model when neither time-constant nor time-varying covariates are present. Van de Pol and Langeheine (1990) also proposed a variant in which the four types of model probabilities could differ across categories of a grouping variable (see also Langeheine and Van de Pol, 2002). A similar model is obtained by replacing the $\mathbf{z}_i$ and $\mathbf{z}_{it}$ in equations (1)–(3) by a single categorical covariate z_i.

3.2 Mixture Markov

The mixture Markov model (Poulsen, 1982) is the special case of the model presented in equations (1)–(3) when there is a single response variable that is assumed to be measured without error. The model is obtained by replacing the more general definition in equation (3) with

$$P(\mathbf{y}_i|w, x_0, x_1, \ldots, x_T, \mathbf{z}_i) = \prod_{t=0}^{T} P(y_{it}|x_t)$$

where $K = M$ and $P(y_{it}|x_t) = 1$ if $x_t = y_{it}$ and 0 otherwise. The product over the multiple response variables and the index j can be omitted because $J = 1$ and y_{it} is assumed not to depend on w and $\mathbf{z}_{it}$ but only on x_t. For this special case the number of latent states (K) is equal to the number of observed states (M) and the relationship between x_t and y_{it} is perfect, which indicates that x_t is measured without error.

A special case of this mixture Markov model is the mover-stayer model (Goodman, 1961). This model assumes that $L = 2$ and that the transition probabilities are fixed to 0 for one class, say for $w = 2$. Members of this class, for which $P(x_t|x_{t-1}, w = 2, \mathbf{z}_{it}) = 1$ if $x_t = x_{t-1}$ and 0 otherwise, are called stayers. Note that the mover-stayer constraint can not only be imposed in the mixture Markov but also in the mixture latent Markov, in which case transitions across imperfectly measured states are assumed not to occur in the stayer class.

Because of the perfect match between x_t and y_{it}, the mixture Markov model can also be defined without latent states x_t; i.e., as:

$$P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L} P(w|\mathbf{z}_i)\, P(y_{i0}|w, \mathbf{z}_i) \times \prod_{t=1}^{T} P(y_{it}|y_{it-1}, w, \mathbf{z}_{it})$$

3.3 Latent Markov model

The latent Markov, latent transition, or hidden Markov model (Baum et al., 1970; Collins and Wugalter, 1992; Van de Pol and De Leeuw, 1996; Vermunt, Langeheine and Böckenholt, 1999; Wiggins, 1973) is the special case of the mixture latent Markov that is obtained by eliminating the time-constant latent variable w from the model, i.e., by assuming that there is no unobserved heterogeneity or that it can be ignored. The latent Markov model can be obtained without modifying the formulae, but by simply assuming that $L = 1$; i.e., that all subjects belong to the same latent class.

The latent Markov model yields estimates for the initial-state and transition probabilities, as well as for how these are affected by covariate values, while correcting for measurement error in the observed states. The model can be applied with a single or with multiple response variables. When applied with a single categorical response variable, one will typically assume that the number of latent states equals the number or categories of the response variable: $K = M$. Moreover, model restrictions are required to obtain an identified model, the most common of which are time-homogeneous transition probabilities or time-homogeneous misclassification probabilities.

When used with multiple indicators, the model is a longitudinal data extension of the standard latent class model (Hagenaars, 1990). The time-specific latent states can be seen as clusters or types which differ in their responses on the J indicators, and the Markovian transition structure is used to describe and predict changes that may occur across adjacent measurement occasions.

3.4 Markov model

By assuming both perfect measurement as in the mixture Markov model and absence of unobserved heterogeneity as in the latent Markov model, one obtains a standard Markov model, which is no longer a latent class model. This model can further serve as a simple starting point for longitudinal applications with a single response variable, where one wishes to assume a Markov structure. It provides a baseline for comparison to the three more extended models discussed above. Use of these more extended models makes sense only if they provide a significantly better description of the data than the simple Markov model.

3.5 Mixture latent growth model

Now we turn to latent class models for longitudinal research that are not transition or Markov models. These mixture growth models are nonparametric random-effects models (Aitkin, 1999; Skrondal and Rabe-Hesketh, 2004; Vermunt and Van Dijk, 2001) for longitudinal data that assume that dependencies between measurement occasions can be captured by the time-constant latent variable w. The most extended variant is the mixture latent growth model, which is obtained from the mixture latent Markov model by imposing the constraint $P(x_t|x_{t-1}, w, \mathbf{z}_{it}) = P(x_t|w, \mathbf{z}_{it})$. This is achieved by replacing equation (2) with

$$P(w, x_0, x_1, \ldots, x_T|\mathbf{z}_i) = P(w|\mathbf{z}_i) \prod_{t=0}^{T} P(x_t|w, \mathbf{z}_{it}).$$

This model is a variant for longitudinal data of the multilevel latent class model proposed by Vermunt (2003): subjects are the higher-level units and time points the lower-level units. It should be noted that application of this very interesting model requires that there be at least two response variables ($J \geq 2$).

In mixture growth models one will typically pay a lot of attention to the modeling of the time dependence of the state occupied at the different time points. The latent class or mixture approach allows identifying subgroups (categories of the time-constant latent variable w) with different change patterns (Nagin, 1999). The extension provided by the mixture latent growth model is that the dynamic dependent variable is itself a (discrete) latent variable which is measured by multiple indicators.

3.6 Mixture growth model

The mixture or latent class growth model (Nagin, 1999; Muthén, 2004; Vermunt, 2006) can be seen as a restricted variant of the mixture latent growth model; i.e., as a model for a single indicator measured without error. The extra constraint is the same as the one used in the mixture Markov model: $K = M$ and $P(y_{it}|x_t) = 1$ if $x_t = y_{it}$ and 0 otherwise.

A more natural way to define the mixture growth model is by omitting the time-varying latent variable x_t from the model specification, as we did for the mixture Markov model. This yields

$$P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L} P(w|\mathbf{z}_i) \prod_{t=0}^{T} P(y_{it}|w, \mathbf{z}_{it})$$

Note that this model is equivalent to a standard latent class model for $T+1$ response variables and with predictors affecting these responses.

3.7 Standard latent class model

When we eliminate both w and the transition structure, we obtain a latent class model that assumes observations are independent across occasions. This is a realistic model only for the analysis of data from repeated cross-sections; i.e., to deal with the situation in which observations from different occasions are independent because each subject provides information for only one time point. One possible way to define this model is

$$P(\mathbf{y}_i|\mathbf{z}_{it_i}) = \sum_{x=1}^{K} P(x|\mathbf{z}_{it_i}) \prod_{j=1}^{J} P(y_{itj}|x, \mathbf{z}_{it_i})$$

where t_i is used to denote the time point for which subject i provides information. This is a standard latent class model with covariates.

4 Application to NYS data

To illustrate the latent class models described above we use data from the nine-wave National Youth Survey (Elliott, Huizinga and Menard, 1989) for which data were collected annually from 1976 to 1980 and at three-year intervals after 1980. At the first measurement occasion, the ages of the 1725 children varied between 11 and 17. To account for the unequal spacing across panel waves and to use age as the time scale, we define a model for 23 time points ($T + 1 = 23$), where $t = 0$ corresponds to age 11 and the last time point to age 33. For each subject, we have observed data for at most 9 time points (the average is 7.93) which means that the other time points are treated as missing values.

We study the change in a dichotomous response variable “drugs” indicating whether young persons used hard drugs during the past year (1=no; 2=yes). It should be noted that among the 11-year-olds in the sample nobody reported to have used hard drugs, which is something that needs to be taken into account in our model specification. Time-varying predictors are age and age squared, and time-constant predictors are gender and ethnicity.

A preliminary analysis showed that there is a clear age-dependence in the reported hard-drugs use which can well be described by a quadratic function: usage first increases with age and subsequently decreases. That is why we used this type of time dependence in all reported models. To give an idea how the time dependence enters in the models, the specific regression model for the latent transition probabilities in the estimated Markov models was:

$$\log \frac{P(x_t = k' | x_{t-1} = k, w, \text{age}_{it})}{P(x_t = k | x_{t-1} = k, w, \text{age}_{it})} = \beta_{0k'k}$$
$$+ \beta_{1k'k} \cdot d_{w=2} + \beta_{2k'k} \cdot \text{age}_{it} + \beta_{3k'k} \cdot (\text{age}_{it})^2$$

where the β coefficients are fixed to 0 for $k' = k$. The variable $d_{w=2}$ is a dummy variable for the second mixture component. For the initial-state, we do not have a model with free parameters but we simply assume that all children start in the no-drugs state at age 11.

In the mixture growth models, we use the following binary logistic regression model for y_{it}:

$$\log \frac{P(y_{it} = 2 | w, \text{age}_{it})}{P(y_{it} = 1 | w, \text{age}_{it})}$$
$$= \beta_{0w} + \beta_{1w} \cdot \text{age}_{it} + \beta_{2w} \cdot (\text{age}_{it})^2$$

where we fix $\beta_{11} = -100$ and $\beta_{21} = \beta_{31} = 0$ to obtain a model in which $w = 1$ represents a non-user class, a class with a zero probability of using drugs at all time points.

Table 23.2 reports the fit measures for the estimated models, where the first set of models do not contain time-constant covariates gender and ethnicity. As can be seen from log-likelihood and BIC values, the various types of Markov models perform much better than the mixture growth models, which indicates that there is a clear autocorrelation structure that is difficult to capture using a growth model. Even with 7 latent classes one does not obtain a fit that is as good as the Markov-type models. Among the Markov models, the most general model – the mixture latent Markov model – performs best. By removing measurement error, simplifying the mixture into a mover-stayer structure, and/or eliminating the mixture structure, the fit deteriorates significantly. The last two models are mixture latent Markov models in which we introduced covariates in the model for the mixture proportions. Both sex and ethnicity seem to be significantly related to the mixture component someone belongs to.

The parameters of the final model consist of the logit coefficients of the model for w, the logit coefficients in the model for the latent transition probabilities, and the probabilities of the measurement model. The latter show that the two latent states are rather strongly connected

Table 23.2 Fit measures for the estimated models with the nine-wave National Youth Survey data set

Model	*Log-likelihood*	*BIC*	*# Parameters*
A. Independence	–5089	10200	3
B. Markov	–4143	8330	6
C. Mixture Markov with $L = 2$	–4020	8108	9
D. Mover-stayer Markov	–4056	8165	7
E. Latent Markov with $K = 2$	–4009	8078	8
F. Mixture latent Markov with $L = 2$ and $K = 2$	–3992	8066	11
G. Mover-stayer latent Markov with $K = 2$	–4000	8068	9
H1. Mixture growth with $L = 2$ ($w = 1$ non-users)	–4381	8792	4
H2. Mixture growth with $L = 3$ ($w = 1$ non-users)	–4199	8457	8
H3. Mixture growth with $L = 4$ ($w = 1$ non-users)	–4113	8315	12
H4. Mixture growth with $L = 5$ ($w = 1$ non-users)	–4077	8273	16
H5. Mixture growth with $L = 6$ ($w = 1$ non-users)	–4037	8223	20
H6. Mixture growth with $L = 7$ ($w = 1$ non-users)	–4024	8227	24
I. F + Gender effect on W	–3992	8066	12
J. F + Gender and Ethnicity effect on W	–3975	8061	15

to the two observed states: $P(y_{it} = 1|x_t = 1) = 0.99$ and $P(y_{it} = 2|x_t = 2) = 0.87$.

The most relevant coefficients in the model for the transition probabilities are the parameters for w. These show that class 2 is the low-risk class, having a much lower probability than class 1 of entering into the use state ($\beta = -2.37; S.E = 0.26$) and a much higher probability of leaving the non-use state ($\beta = 3.72; S.E. = 0.68$). Combining these estimates with the quadratic time dependence of the transitions yields a probability of moving from the non-use to the use state equal to 2.8% at age 12, 23.4% at age 21, and 0.6% at age 33 for $w = 1$, and equal to 0.3% at age 12, 2.8% at age 21, and 0.1% at age 33 for $w = 2$. The probability of a transition from the use to the non-use state equals 0.1% at age 12, 20.5% at age 26, and 6.2% for $w = 1$, and 4.1% at age 12, 91.4% at age 26, and 73.1% at age 33 for $w = 2$.

The parameters in the logistic regression model for w shows that males are less likely to be in the low-risk class than females ($\gamma = -0.58; S.E. = 0.14$) and that blacks are more likely to be in the low-risk class than whites ($\gamma = 0.79; S.E = 0.22$). Hispanics are less likely ($\gamma = -0.46; S.E. = 0.33$) and other ethnic groups more likely ($\gamma = 0.25; S.E = 0.52$) to be in class 2 than whites, but these effects are non-significant.

5 Discussion

We presented a general framework for the analysis of discrete-time longitudinal data and illustrated it with an empirical example in which the Markov-like models turned out to perform better than the growth models.

The approach presented here can be expanded in various ways. First, while we focused on models for categorical response variables, it is straightforward to apply most of these models to variables of other scale types, such as continuous dependent variables or counts. Other extensions include the definition of multiple processes with multiple x_t or of higher-order Markov processes. Models that are getting increased attention are those that combine discrete and continuous latent variables. Finally, the approach can be expanded to deal with multilevel longitudinal data, as well as with data obtained from complex survey samples. Each of these extensions is implemented in the Latent GOLD software that we used for parameter estimation.

Appendix A: Baum-Welch algorithm for the mixture latent Markov model

Maximum likelihood (ML) estimation of the parameters of the mixture latent Markov model involves maximizing the log-likelihood function:

$$L = \sum_{i=1}^{N} \log P(\mathbf{y}_i|\mathbf{z}_i)$$

a problem that can be solved by means of the EM algorithm (Dempster, Laird and Rubin, 1977). In the E step, we compute

$$P(w, x_0, x_1, \ldots, x_T|\mathbf{y}_i, \mathbf{z}_i) = \frac{P(w, x_0, x_1, \ldots, x_T, \mathbf{y}_i|\mathbf{z}_i)}{P(\mathbf{y}_i|\mathbf{z}_i)}$$

which is the joint conditional distribution of the $T + 2$ latent variables given the data and the model parameters. In the M step, one updates the model parameters using standard ML methods for logistic regression analysis and using an expanded data matrix with $P(w, x_0, x_1, \ldots, x_T|\mathbf{y}_i, \mathbf{z}_i)$ as weights.

It should be noted that in a standard EM algorithm, at each iteration, one needs to compute and store the $L \cdot K^{T+1}$ entries of $P(w, x_0, x_1, \ldots, x_T|\mathbf{y}_i, \mathbf{z}_i)$ for each subject or, with grouped data, for each unique data pattern. This implies that computation time and computer storage increases exponentially with the number of time points, which makes this algorithm impractical or even impossible to apply with more than a few time points (Vermunt, Langeheine and Böckenholt, 1999). However, because of the collapsibility of the mixture latent Markov model, it turns out that in the M step of the EM algorithm one needs only the marginal distributions $P(w|\mathbf{y}_i, \mathbf{z}_i)$, $P(w, x_t|\mathbf{y}_i, \mathbf{z}_i)$, and $P(w, x_{t-1}, x_t|\mathbf{y}_i, \mathbf{z}_i)$. The Baum-Welch or forward-backward algorithm obtains these quantities directly rather than first computing $P(w, x_0, x_1, \ldots, x_T|\mathbf{y}_i, \mathbf{z}_i)$ and subsequently collapsing over the remaining dimensions as would be done in a standard EM algorithm (Baum et al., 1970; McDonald and Zucchini, 1997). This yields an algorithm that makes the mixture latent Markov model applicable with any number of time points. Whereas the original forward-backward algorithm was for latent (hidden) Markov models without covariates and a single response variable, here we provide a generalization to the more general case with a mixture w, covariates $\mathbf{z}_i$, and multiple responses.

The two key components of the Baum-Welch algorithm are the forward probabilities α_{iwx_t} and the backward probabilities β_{iwx_t}. Because of our generalization to the mixture case, we need an additional quantity γ_{iw}. These three quantities are defined as follows:

$$\alpha_{iwx_t} = P(x_t, \mathbf{y}_{i0} \ldots \mathbf{y}_{it}|w, \mathbf{z}_i),$$
$$\beta_{iwx_t} = P(\mathbf{y}_{i(t+1)} \ldots \mathbf{y}_{iT}|x_t, w, \mathbf{z}_i),$$
$$\gamma_{iw} = P(w, \mathbf{y}_i|\mathbf{z}_i).$$

Using α_{iwx_t}, β_{iwx_t}, and γ_{iw}, one can obtain the relevant marginal posteriors as follows:

$$P(w|\mathbf{y}_i, \mathbf{z}_i) = \frac{\gamma_{iw}}{P(\mathbf{y}_i|\mathbf{z}_i)}, \tag{4}$$

$$P(w, x_t|\mathbf{y}_i, \mathbf{z}_i) = \frac{\alpha_{iwx_t}\beta_{iwx_t}}{P(\mathbf{y}_i|\mathbf{z}_i)}, \tag{5}$$

$$P(w, x_{t-1}, x_{t-1}, w|\mathbf{y}_i, \mathbf{z}_i) = \frac{\gamma_{iw}\alpha_{iwx_{t-1}} P(x_t|x_{t-1}, w, \mathbf{z}_{it}) P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it}) \beta_{iwx_t}}{P(\mathbf{y}_i|\mathbf{z}_i)}, \tag{6}$$

where $P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L} \gamma_{iw}$, and $P(x_t|x_{t-1}, w, \mathbf{z}_{it})$ and $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})$ are model probabilities.

The key element of the forward-backward algorithm is that $T+1$ sets of α_{iwx_t} and β_{iwx_t} terms are computed using recursive schemes. The forward recursion scheme for α_{iwx_t} is:

$$\alpha_{iwx_0} = P(x_0|w, \mathbf{z}_{i0}) P(y_{i0}|x_0, w, \mathbf{z}_{i0}),$$
$$\alpha_{iwx_t} = \left\{ \sum_{x_{t-1}=1}^{K} \alpha_{iwx_{t-1}} P(x_t|x_{t-1}, w, \mathbf{z}_{it}) \right\} \times P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})$$

for $t = 1$ up to $t = T$. The backward recursion scheme for β_{iwx_t} is:

$$\beta_{iwx_T} = 1,$$
$$\beta_{iwx_t} = \sum_{x_{t+1}=1}^{K} \beta_{iwx_{t+1}} P(x_{t+1}|x_t, w, \mathbf{z}_{it}) \times P(\mathbf{y}_{it+1}|x_{t+1}, w, \mathbf{z}_{it})$$

for $T-1$ down to $t = 0$. The quantity γ_{iw} is obtained as:

$$\gamma_{iw} = \sum_{x_t=1}^{K} P(w|\mathbf{z}_i)\, \alpha_{iwx_t} \beta_{iwx_t},$$

for any t. So, first we obtain α_{iwx_t} and β_{iwx_t} for each time point and subsequently we obtain γ_{iw}. Next, we compute $P(w|\mathbf{y}_i, \mathbf{z}_i)$, $P(w, x_t|\mathbf{y}_i, \mathbf{z}_i)$, and $P(w, x_{t-1}, x_t|\mathbf{y}_i, \mathbf{z}_i)$ using equations (4), (5), and (6). In the M step, these quantities are used to obtain new estimates for the mixture latent Markov model probabilities appearing in equations (2) and (3) using standard methods for logistic regression analysis.

The only change required in the above formulas when there is missing data is that $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})$ is replaced by $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})^{\delta_{it}}$ in each of the above equations, where $\delta_{it} = 1$ if $\mathbf{y}_{it}$ is observed and 0 if $\mathbf{y}_{it}$ is missing. This implies that $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})$ is "skipped" when $\mathbf{y}_{it}$ is missing. In the M step, cases with missing responses at occasion *t* do not contribute to the estimation of the response probabilities for that occasion, but they do contribute to the estimation of the other model probabilities.

Appendix B: Examples of Latent GOLD syntax files

The Latent GOLD 5.0 software package (Vermunt and Magidson, 2005) implements the framework described in this article. In this appendix, we provide examples of input files for estimation of mixture latent Markov models, mixture Markov, latent Markov, and mixture growth models.

The data should be in the format of a person-period file, where for the Markov-type models periods with missing values should also be included in the file since each next record for the same subject is assumed to be the next time point. The definition of a model contains three main sections: "`options`", "`variables`" and "`equations`".

An example of the most extended model, the mixture latent Markov model is the following:

```
options
   missing=all;
   coding=first;
variables
   caseid id;
   dependent drugs nominal;
   independent gender nominal,
   ethnicity nominal, age numeric,
   age2 numeric;
   latent
     W nominal 2;
     X nominal markov 2;
equations
   W < - 1 + gender + ethnicity;
   X[=0] <- (-100) 1;
   X <- (a~) 1 | X[-1] + (b~) W | X[-1]+
   (c~) age | X[-1] + (d~) age2|X[-1];
   drugs <- (e~) 1 | X;
```

In the `options` section, only the two commands for which we changed the default setting is shown. The statement "`missing=all`" indicates that all records with missing values should be retained in the analysis. The option "`coding=first`" requests dummy coding for the nominal variables using the first category as the reference category.

In the `variables` section we define the caseid variable connecting the multiple records of one person, the latent, dependent (or response) and independent variables to be used in the analysis, as well as various attributes of these variables, such as their scale type and, for categorical latent variables, their number of categories and whether they vary over time (indicated with the statement `markov`).

The `equation` section contains four equations: one for the mixture variable (`W`), one for the initial state (`X[=0]`), one for the state at time point *t* (`X`) conditional on the state at $t - 1$ (`X[-1]`), and one for the response variable. With more response variables, one would have a separate equation for each response variable. The logit model for `W` contains an intercept (the term "1") and effects of gender and ethnicity. The parameter labels, `a`, `b`, `c`, `d`, and `e` are given in parentheses. The model for `X[=0]` contains an intercept that is fixed to –100, which means that everyone starts in latent state 1. The model for `X` is parameterized in such a way that the intercept and the effects of `W`, `age`, and `age2` can be interpreted as effects on the logit of a transition (as in the equation provided in the text). This is achieved by the conditioning "`| X[-1]`" combined with the tilde "~" in the parameter label, which yields a special coding of logit coefficients in which the no-change category serves as the reference category. The model for the response variable `drugs` contains an intercept which varies across latent states, with the same type of coding as used for the transition.

A mixture Markov is obtained with the extra line "`e = -100;`". This fixes the logit parameters in the model for the response variable to −100, which because of the special coding induced by the tilde in the parameter label yields a perfect relationship between `X` and `drugs`. The 2-class mixture can be changed into a mover-stayer structure with the additional line "`b = -100;`" which fixes the transition probabilities to 0 for the second class. This restriction can be used in the mixture Markov and in the mixture latent Markov model. A latent Markov model is obtained either by removing `W` from the `variables` and `equations` sections or by setting its number of categories to 1.

A mixture growth model is obtained by removing `X` from the variables section and replacing the `equations` section with the following:

```
equations
    W <- 1 + gender + ethnicity;
    drugs <- (a~) 1 | W + (b~) age | W +
    (c~) age2 | W;
    a[1] = -100;
    b[1] = 0;
    c[1] = 0;
```

The constraint on the intercept indicates that the first mixture component does not use drugs with probability 1. The other two constraints fix the redundant `age` and `age2` effects for class one equal to 0.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Aitkin (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55: 218–234.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41: 164–171.

Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions. *Psychological Methods*, 10: 65–82.

Collins, L.M. and Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27: 131–157.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39: 1–38.

Elliott, D.S., Huizinga, D. and Menard, S. (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*. New York: Springer-Verlag.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. London: Chapman & Hall.

Fay, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81: 354–365.

Goodman, L.A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56: 841–868.

Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I – A modified latent structure approach. *American Journal of Sociology*, 79: 1179–1259.

Hagenaars, J.A. (1990). *Categorical Longitudinal Data—Loglinear Analysis of Panel, Trend and Cohort Data*. Newbury Park: Sage.

Langeheine, R. and Van de Pol, F. (2002). Latent Markov chains. In J.A. Hagenaars and A.L. McCutcheon (eds), *Applied Latent Class Analysis*, pp. 304–341. Cambridge, UK: Cambridge University Press.

Little, R.J. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

McDonald, I.L. and Zucchini, W. (1997), *Hidden Markov and Other Models for Discrete Valued Time Series*. London: Chapman & Hall.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Menard, S, (2002). *Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage.

Muthén, B. (2004). Latent variable analysis. Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Chapter 19, pp. 345–368. Thousand Oaks, CA: Sage Publications.

Nagin, D.S. (1999). Analyzing developmental trajectories: a semiparametric group-based approach. *Psychological Methods*, 4: 139–157.

Poulsen, C.S. (1982). *Latent Structure Analysis with Choice Modeling Applications*. Arhus: the Arhus School of Business Administration and Economics.

Schafer, J.L. (1997). *Statistical Analysis with Incomplete Data*. London: Chapman & Hall.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. London: Chapman & Hall/CRC.

Van de Pol, F. and De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, 15: 118–141.

Van de Pol, F. and Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, 213–247.

Vermunt, J.K. (1997). *Log-linear Models for Event Histories*. Thousand Oaks, CA: Sage.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33: 213–239.

Vermunt, J.K. (2006). Growth models for categorical response variables: Standard, latent-class, and hybrid approaches. In K. van Montfort, H. Oud and A. Satorra (eds), *Longitudinal Models in the Behavioral and Related Sciences*. Wahwah, NJ: Erlbaum.

Vermunt, J.K., Langeheine, R. and Böckenholt, U. (1999). Latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24: 178–205.

Vermunt, J.K. and Magidson, J. (2005). *Technical Guide to Latent GOLD 4.0: Basic and Advanced*. Belmont, Massachusetts: Statistical Innovations.

Vermunt, J.K. and Van Dijk, L. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modelling Newsletter*, 13: 6–13.

Wiggins, L.M. (1973). *Panel analysis.* Amsterdam: Elsevier.

This page intentionally left blank

Part V

Timing of Qualitative Change: Event History Analysis

This page intentionally left blank

Chapter 24

Nonparametric methods for event history data: descriptive measures

C. M. Suchindran

1 Introduction

An individual's life history is often characterized by a history of transitions to a number of life events. Occurrence of a specific event is considered as a transition to a particular state. These states can include both transient (states from which transitions are possible) and absorbing states (states from which escape is not possible). Suppose that each subject begins life in one of several states (usually in a transient state) and that at each point in time will either be in the same state or will make a transition to one of the possible states. Event history data for a subject can be described by a set of events experienced by the subject and the timing of these transitions. Thus event history data usually consists of timing of occurrence of multiple events. The data can have additional complexities with some durations in some states being censored. Also, it is possible that studies by design observe only partial history for each subject (we will describe below a number of such designs).

To illustrate the data and describe notations we will look at the following example describing a woman's marital history. A married woman can divorce, be widowed or die. Therefore, the life history will consist of three transient states (married, divorced, and widowed) and an absorbing state (death). The data will record age at which the subject experiences a specific event along with the type of event. Without any loss of generality, denote states 1, 2, S_1 as transient states and states $S_1 + 1 \ldots$ S as absorbing. In the example above, denote state 1 as married, state 2 as widowed, state 3 as divorced, and state 4 as death. In this case $S_1 = 3$ and S = 4. Let z_n denote the state corresponding to a subjects nth transition and let z_0 denote the initial state. In the example above all women start the observation at their first marriage and therefore $z_0 = 1$ (married). If T_n represents the time at which transitions occur, for a subject i the event history data vector will be of the form:

$$H_i = \{z_0, T_1, z_1, \ldots, T_m, z_m\}, \text{ where } 1 \leq z_j \leq S_1$$
$$\text{for } j < m \text{ and } S_1 < z_m \leq S$$

Suppose a married woman ends her marriage by divorce at duration T_1 (time measured from first marriage) and dies at duration T_4, her data vector will be of the form:

$$H_i = \{z_0 = 1, z_1 = 3, T_1, z_4 = 4, T_4\}$$

Note that, in this case, the event history ends when an absorbing state (death) is entered. In most observational settings, some event histories will be incomplete due to right censoring because the subject has not reached an absorbing state by the time the subject was last observed. In this case the notation (1) can still be used by taking $z_m = S+1$ to denote the time T_m as the censored time.

Most often studies by design collect only a partial event history. The usual single-event survival data consist of two states (one transient and one absorbing state) and time to transition to an absorbing state is noted. Usually right censoring occurs in the data. In the marital history example, suppose an investigator is interested in time to marital disruption of first marriage and treats widowhood and divorce together as an absorbing event signifying marital disruption. For simplicity in this situation we will assume that death of the subject or the end of observation period without the occurrence of marital disruption or death will be treated as a censored observation. For presentation in this paper, such data will be called *single spell data with censoring*. Unlike single spell data with one absorbing state, the data can have one transient state and more than one absorbing state. For example, in the study of marital history, the investigator may want to examine whether or not the marriage ended by widowhood or divorce. Once again the data can be censored by death of the subject or by the end of observation while still married. We will call this data, *single spell data with competing risks*. When the data involves more than one transient state and absorbing state, we will label it as *multistate data*. When a subject experiences an event of the same type (transition among two transient states) repeatedly over time, we call the recorded data *recurrent event data*.

In several practical situations, data observation will be further limited. For example, in a single spell examination, it is possible that the collected information consists of an observation time and the knowledge that the event of interest occurred before the observation time. In this situation, the exact time occurrence of the event (transition) will not be known. A similar situation can occur in the collection of recurrent event data where we know the observation time and the number events that occurred before the observation time. In this case the subject is examined at only one point in time and the exact times of transitions are unknown. Such data are called *current status data*. Occasionally in recurrent event situations, the data consists of the observation time and the time elapsed between the last occurrence of the event and the observation time. Such data are referred as *backward recurrence time data.*

In this paper we will describe various nonparametric methods to analyze event history data. The method will depend on the type of available data as described earlier. In Section 2 analysis of single spell data with censoring will be introduced. The analysis will be extended to include competing risks in Section 3. The multistate data description is introduced next. Although recurrent event data can be considered as a special case of multistate data, methods specific to the analysis of recurrent events appear in the literature. We will review them in Section 5. Following discussion of recurrent event data, the analysis of current status and backward recurrence data will be introduced in Section 6. The emphasis in this paper is on nonparametric measures with the goal to describe the data. Therefore measures based on parametric models will not be discussed, and the paper will not discuss introduction of covariates in the data. These issues will be addressed in subsequent chapters.

2 Single spell data with censoring

As mentioned earlier single spell data examines the transition from a transient state (e.g., alive) to an absorbing state (death). The data consists of time to a specified event from the beginning of exposure period. If at the time of observation the event has not occurred, the information for this case is considered as right censored and the time is measured as the time elapsed from the time of exposure to the time of observation. The main summary measure of this data is a nonparametric estimate of the survival function. Kaplan and Meier (1958) provide an estimate of the survival function when censoring is present in the data. Let $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ denote the distinct ordered times of event (not counting censoring times). Let d_i be the number of events at $t_{(i)}$, and let n_i be the number not experiencing the event just before $t_{(i)}$ (called number exposed to risk at time $t_{(i)}$). Then the Kaplan-Meier estimator of the survival function is

$$\hat{S}(t) = \prod_{i:t_{(i)}<t} \left(1 - \frac{d_i}{n_i}\right) \tag{1}$$

This estimator is also known as the product limit estimator. The variance of the estimated survival function is obtained by using the Greenwood formula

$$Var(\hat{S}(t_i)) = [\hat{S}(t_i)]^2 \sum_{j=1}^{i} \frac{d_j}{n_j(n_j - d_j)} \tag{2}$$

The data can be summarized by several other important functions of the survival function. The main summarizing function is the hazard function h(t) which expresses the relative rate of change in the survival function. Thus the hazards function is

$$h(t) = -\frac{S'(t)}{S(t)} \tag{3}$$

The function $\Lambda(t) = \int_0^t h(\tau)d\tau$ is called the cumulative hazard function which has the relation to the survival function

$$\Lambda(t) = -\log S(t)$$

Thus a cumulative hazard function can be estimated by simply computing the negative of the log of the estimate of the survival function. It has an interpretation as the expected number of events in (0, t] per unit at risk of experiencing the event. An alternative approach to estimate the cumulative hazard function directly using the Nelson-Aalen estimator

$$\hat{\Lambda}(t_{(i)}) = \sum_{j=1}^{i} \frac{d_j}{n_j} \tag{4}$$

This estimator is sometimes preferred because of its strong theoretical justification. Breslow (1972) suggested estimating the survival function as

$$\hat{S}(t) = \exp\{-\hat{\Lambda}(t)\} \tag{5}$$

The proportion of individuals experiencing the event or the cumulative probability of experiencing the event by time t, denoted as $\hat{F}(t)$ is calculated as

$$\hat{F}(t) = 1 - \hat{S}(t)$$

Quantile measures derived from the survival function are also used to summarize the data. For example, median time to event (t_m) is calculated from the relation

$$S(t_m) = \frac{1}{2}$$

2.1 Example

The data for this illustration is taken from a demographic survey. The event of interest is the occurrence of the fourth birth. The exposure period starts at the time of the third birth. For those who had a fourth birth, time is calculated as the time elapsed between the third and fourth birth. Those who did not experience a fourth birth at the time of the survey are considered as censored observations. For them the time is calculated as the time elapsed between their third birth and the survey date. There were 303 women in the data set. Table 24.1 gives a tabulation of the data.

The survival functions and the corresponding standard errors are calculated using equations (1)–(5). The survival functions calculated by the Kaplan-Meier and Nelson-Aalen

Table 24.1 Time to fourth order birth

Time in years	*Number in the risk set n_i*	*Number of events d_i*	*Kaplan Meier $\hat{S}(t)$*	*Standard error $(\hat{S}(t))$*	*Nelson-Aalen cumulative hazard $\hat{\Lambda}(t)$*	*Nelson-Aalen survival function $\hat{\hat{S}}(t)$*	*Standard error of $\hat{\hat{S}}(t)$*
1	303	22	0.9274	0.0149	0.0726	0.9300	0.0144
2	249	56	0.7188	0.0271	0.2975	0.7427	0.0251
3	169	40	0.5487	0.0313	0.5342	0.5861	0.0295
4	116	18	0.4635	0.0323	0.6894	0.5019	0.0313
5	80	4	0.4404	0.0327	0.7394	0.4774	0.0320
6	69	3	0.4212	0.0331	0.7828	0.4571	0.0328
7	59	8	0.3641	0.0342	0.9184	0.3992	0.0344
8	47	4	0.3331	0.0346	1.0035	0.3666	0.0353
9	37	3	0.3061	0.0351	1.0846	0.3380	0.0361
10	27	4	0.2608	0.0365	1.2328	0.2915	0.0379
11	19	2	0.2333	0.0375	1.3380	0.2624	0.0393
12	14	3	0.1833	0.0390	1.5523	0.2118	0.0411
13	6	1	0.1528	0.0428	1.7190	0.1792	0.0459

methods are usually quite close. This is particularly true when the number of events is small relative to the number in the risk set. The table shows that nearly 15% of the women with a third birth did not have a fourth birth (Kaplan-Meier estimate) in 13 years. The median time to fourth birth is 3.57 years.

So far we have described a method to analyze right censored single spell data collected in single years. Sometimes the data record event times in intervals. Assuming censoring occurs uniformly in the interval, the exposure time for those who are censored in the interval is given as half the length of the interval. When left censoring is present in the data minor adjustments can also be made to the calculations of the risk set in calculating the survival function (Guo, 1993).

3 Single spell data with competing risks

Earlier we examined methods to analyze single spell data with no competing event present. For example, in the example above we have examined the occurrence of fourth birth in the absence of marital disruption or death of the individual. Assume that there are k competing events in the population. Chiang (1968) introduces the following quantities. Suppose events occur at distinct ordered times $t_{(1)} = 0 < t_{(2)} < \ldots < t_{(n)}$. The data will note the time and the type of event. Define $Q_{t\delta}$ as the crude probability of occurrence of an event of type R_δ, $\delta = 1, 2, \ldots$ k at time t in the presence of other competing risks. Note that $\sum\limits_{\delta=1}^{k} Q_{t\delta} = q_t$, the probability of occurrence of an event, regardless of type at time t. The goal of the analysis is to compute the probability that an individual experiences a specific event by a given duration t. Using the product limit estimator we calculate the cumulative probability of not experiencing any event by time t as

$$\hat{S}(t) = \prod_{i=0}^{t-1} [1 - q_i]$$

The crude probability $Q_{t\delta}$ is calculated as

$$Q_{i\delta} = \frac{d_{i_\delta}}{n_i} \tag{6}$$

where $d_{i\delta}$ is the number of events of type R_i at time t_i and n_i is the number in the risk set at that time. The cumulative probability that an individual experiences a specific event R_δ, denoted as $CQ_{t\delta}$, is calculated as

$$CQ_{t\delta} = \sum_{i=0}^{t-1} S(i) Q_{i\delta} \tag{7}$$

Variance of the estimated cumulative probability is calculated as follows:

Denote $\hat{p}_i = 1 - \hat{q}_i$. $Variance(\hat{p}_i) = \hat{p}_i(1 - \hat{p})/n_i$, where n_i is the number at risk at time t_i.

Also $Variance(\hat{Q}_{i\delta}) = \hat{Q}_{i\delta}(1 - \hat{Q}_{i\delta})/n_i$ and $Covariance(\hat{p}_i \hat{Q}_{i\delta}) = -\hat{p}_i \hat{Q}_{i\delta}/n_i$.

Let V_t be the variance of the cumulative probability $C\hat{Q}_{t\delta}$. This variance can be computed using the following formula

$$V_t = \sum_{j=0}^{t-1} A_{jt}^2 Var(\hat{p}_j) + \sum_{j=0}^{t} B_j^2 Var(\hat{Q}_{j\delta}) + \sum_{j=0}^{t} A_{jt} B_j cov(\hat{p}_j \hat{Q}_{j\delta}) \tag{8}$$

where A_{jt} follows the recurrence relation $A_{jt} = A_{jt-1} + S(t)Q_{t\delta}$, $B_j = \hat{S}(t_j) \quad (B_0 = 1)$ and $A_{jt} = \sum_{i=j}^{t} \hat{S}(t_i)\hat{Q}_{i\delta}$.

3.1 Example

A demographic survey collected marital history from 17,045 women. The goal is to examine the disruption of the first marriage. The two competing causes of marital disruption are divorce and widowhood. Women who are in an intact marriage at the time of survey are considered as censored observations. For them time is calculated as the time from first marriage to the survey. Otherwise the duration of marriage at the time of disruption is noted. Table 24.2 shows a partial tabulation of data.

Equations (6) and (7) are used to obtain crude probabilities of divorce and widowhood and are presented in Table 24.3. The table shows that the probability of not experiencing marital disruption is 0.90360 before 7 years and the probability that the marriage will end in 7 years is 1–0.90360 = 0.09640. The table also shows that the probability that the marriage will end due to death of spouse in seven years is 0.00315 and due to divorce is 0.09325. The standard errors calculated using equation (8) are included in Table 24.3.

4 Multistate data

So far we have examined event history data that involves one transient state and one or more absorbing states. In this section we will examine

Table 24.2 Marital disruptions by divorce and widowhood

Duration of marriage in completed years	*Risk set*	*Number divorced*	*Number widowed*	*Number of marital disruptions*	*Number still married at survey*
0	17045	140	1	141	88
1	16816	211	3	214	222
2	16380	272	2	274	523
3	15583	256	8	264	405
4	14914	232	12	244	452
5	14218	193	10	203	555
6	13460	193	13	206	539
7	12715	174	12	186	543

Table 24.3 Cumulative probabilities and standard errors

Duration of marriage in completed years	q_x	$S(x)$	Q_{xw}	Q_{xd}	*Cumulative probability of widowhood* CQ_{widow}	*Cumulative probability of divorce* $CQ_{divorce}$	*Standard error* $C\hat{Q}_{widow}$	*Standard error* $C\hat{Q}_{Divorce}$
0	0.00827	1.00000	0.00006	0.00821	0.00006	0.00821	0.000059	0.00069
1	0.01273	0.99173	0.00018	0.01255	0.00024	0.02065	0.000118	0.00109
2	0.01673	0.97911	0.00012	0.01661	0.00036	0.03691	0.000145	0.00145
3	0.01694	0.96273	0.00051	0.01643	0.00085	0.05273	0.000227	0.00173
4	0.01636	0.94642	0.00080	0.01556	0.00160	0.06745	0.000315	0.00196
5	0.01428	0.93093	0.00070	0.01357	0.00226	0.08009	0.000377	0.00213
6	0.01530	0.91764	0.00097	0.01434	0.00315	0.09325	0.000450	0.00231
7		0.90360						

processes that can take multiple states that include more than one transient state. Detailed description of multistate data analysis can be found in Namboodiri and Suchindran (1987). Suppose there are a finite number of states with two or more transient states that individuals can move in and out at various time points. The event history data, possibly censored, records the time at which various transitions occur. For example, migration history data records an individual's movements in life until death or until a point of observation at which the history is censored. Let $X(t)$ denote the state occupied by an individual at time t and there are K states of which K_1 states are transient and K_2 states are absorbing $(K_1 + K_2 = K)$. Under Markov assumptions, the process is governed by a set of *transition probabilities* that the state occupied at time t is j given that the state occupied at time s $(0 \leq s \leq t)$ is i, denoted as:

$$q_{ij}(s,t) = P[X(t) = j|X(s) = i] \tag{9}$$

Note that $\sum_{j=1}^{K} q_{ij}(s,t) = 1$ and if i is an absorbing state $q_{ij}(s,t) = 1$ for $j = i$ and zero otherwise. Because X (t) changes continuously with time, the process is also sometimes described in terms of transition intensity defined as:

$$r_{ij}(s) = \lim_{u \to 0} \frac{q_{ij}(s, s+u)}{u}, \quad i \neq j \tag{10}$$

$$\text{and} \quad r_{ii}(s) = \lim_{u \to 0} \frac{[q_{ii}(s, s+u) - 1]}{u},$$

$$\text{so that} \quad \sum_{j=1}^{K} r_{ij}(s) = 0$$

The transition probabilities are usually put in a matrix $Q(s,t)$ with its $(ij)^{th}$ element being $q_{ij}(s,t)$ and similarly form a matrix of transition intensities $R(s)$. Several summary measures of the process can be computed.

State occupancy probabilities provide the probability of being in a particular state after a longer period given the initial state. Formally, it is the probability that an individual is in state j at time t, given that an individual occupies state i at time s. Assuming a partitioning of the interval (s, t) as $s = t_0 < t_1 < \ldots\ldots < t_n = t$, then the state occupancy probabilities are the elements of the matrix

$$Q(s,t) = \prod_{h=0}^{n-1} Q(t_h, t_{h+1}) \tag{11}$$

A second useful measure is the expected length of time spent in a specified state. Denote $e_{ij}(s,t)$ as the time spent in state j during the interval (s, t) for an individual in state i at time s. Form the corresponding matrix $E(s,t) = \{e_{ij}(s,t)\}$. Then

$$E(s,t) = \int_{s}^{t} Q(s,\tau)d\tau \tag{12}$$

A third useful measure is the expected number of visits to specific states in a specified time interval. Let $m_{ij}(s,t)$ denote the number of visits to state j in the interval (s, t) for an individual in state i at time s. Form the corresponding matrix M(s, t). Then

$$M(s,t) = \int_{s}^{t} Q(s,u)B(u)du \tag{13}$$

where B(u) is obtained from R(u) by replacing its diagonal elements with zero. Event history data allows one to compute transition probabilities and the corresponding summary measures.

For convenience tabulate data in small intervals $(t, t+h)$. Let $d_{ij}(t, t+h)$ denote the number of transitions from state i to state j in the interval. Also denote $n_i(t)$ as the number of individuals in state i at time t and $c_{ij}(t, t+h)$ the number of individuals in state i at time t who were censored in the interval $(t,\ t+h)$. Then the transition probability can be estimated as:

$$\hat{q}_{ij}(t,t+h) = \frac{d_{ij}(t,t+h)}{n_t - \frac{1}{2}c_i(t,t+h)} \tag{14}$$

Use these estimated transition probabilities to from the matrix $\hat{Q}(t, t+h)$. The state occupancy probabilities at time t are calculated as $\prod_{h=0}^{n-1} Q(t_h, t_{h+a})$.

4.1 Example

The following data on contraceptive use was created to illustrate the method. In this setup the multistate process has three states: use, non-use and pregnancy. The use and non-use states are considered as transient because women can move back and forth from these states. For illustration, we treat pregnancy in this example state as an absorbing state. The dataset contained contraceptive history of 1835 women among whom 1765 accepted contraceptive at the time of recruitment. The remaining 70 women started in non-use state. The following table shows the transition in the first month:

	Destination state			
Origin		Non-		
state	Use	use	Pregnant	Censored
Use	1721	23	18	3
Non-use	27	39	3	1

Equation (14) was used to calculate the transition probability matrix for month 1: $Q(0,1) =$

	Destination state		
Origin state	Use	Non-use	Pregnant
Use	0.9881	0.0102	0.0017
Non-use	0.3885	0.5683	0.0432
Pregnant	0	0	1

Transition data during month 2 (labeled as 1–2) were tabulated as:

	Destination state			
Origin		Non-		
state	Use	use	Pregnant	Censored
Use	1703	27	18	0
Non-use	13	45	3	1

The corresponding transition matrix is calculated as: $Q(1,2) =$

	Destination state		
Origin state	Use	Non-use	Pregnant
Use	0.9742	0.0154	0.0103
Non-use	0.2113	0.7398	0.0487
Pregnant	0	0	1

The contraceptive status at the end of month 2 is calculated using equation (11) as $Q(0,2) = Q(0,1) * Q(1,2)$. The resulting matrix is:

	Destination state (at the end of month 2)		
Origin state (at time zero)	Use	Non-use	Pregnant
Use	0.9648	0.0228	0.0124
Non-use	0.4986	0.4265	0.0749
Pregnant	0	0	1

The results show that the probability that a non-user at the beginning of the study will be in the use status at the end of month two is 0.4968 and in a pregnant state is 0.0749. Repeating the calculations to the end of six months (data not shown):

	Destination state (at the end of month 6)		
Origin state (at time zero)	Use	Non-use	Pregnant
Use	0.8606	0.0789	0.0605
Non-use	0.6012	0.2369	0.1619
Pregnant	0	0	1

Duration of contraceptive use for those who are in use status or non-use status can be calculated using equation 12. When data are tabulated in one unit (one month in the example here) intervals the integral can be approximated by sum as $E(0,t) = \sum_{i=0}^{n-1} Q(t_i, t_{i+1})$. A simple linear approximation can be to set the interval length to one if the interval length is more than one unit. Calculations based on the data show that the average length of stay in various contraceptive states at the end of six months is as follows:

		Average duration (in months) in state	
		Use	Non-use
Initial state	Use	5.52	0.30
	Non-use	3.14	2.21

The data shows that a woman in the non-use state at the beginning of the study will spend only 2.21 months in the non-use state during the first six months of observation.

A third summary measure is to examine the average number of visits to each state during a fixed time period of observation for a woman starting in a specific state. For this purpose we will use equation (13). In order to calculate the B matrix in equation (13) we will use the following relationship between the transition matrix (Q) and the transition intensity matrix R. Let $Q_{11}(s,t)$ denote the submatrix indicating transitions among the absorbing state with corresponding transition intensity matrix for the interval $R_{11}(u)$ for u in the interval (s, t). Then

$$\begin{aligned} R_{11} = -\frac{1}{h}[(I - Q_{11}(s,s+h)) \\ +(I - Q_{11}(s,s+h))^2/2 \\ +(I - Q_{11}(s,s+h))^3/3 + \ldots] \quad (15) \end{aligned}$$

In the example above: $Q_{11}(0,1) =$

		Destination state	
		Use	Non-use
Origin state	Use	0.9881	0.0102
	Non-use	0.3884	0.5683

Using equation (15) the resulting R_{11} matrix for the interval is:

		Destination state	
		Use	Non-use
Origin state	Use	−0.01470	0.0134
	Non-use	0.5083	−0.5640

The B matrix is obtained by setting the diagonal elements of R_{11} matrix as zero.

In the example here the integral in equation 13 is approximated as:

$$M(0,6) = \sum CQ_{11}(0,t_i)B(t_i)$$

The results show the average number of visits specific for each initial state as follows:

$M_{11}(0,6) =$

		Destination state	
		Use	Non-use
Origin state	Use	0.0161	0.0814
	Non-use	0.7430	0.0362

The result can be interpreted as follows. A cohort of 1000 women who at the beginning of observation are in the use state will, on an average, make 81 transitions to the non-use state and 16 revisits to the user state in six months. Similarly, 1000 women starting in the non-use state will, on an average, make 743 visits to the user state and 36 revisits to the non-use state.

5 Recurrent event data

Recurrent event data is generated when a subject experiences a specific event several times during the period of observation. One feature of this data is that event times, possibly censored, are ordered and correlated. Such data are frequently observed in longitudinal follow-up studies. Event history recorded through retrospective studies also generates recurrent event data. For example, several demographic surveys collect timing of child birth to women or migration events through retrospective enquiries. In this data the event history will be censored at the duration at the time of survey. Unlike in retrospective survey data, in a prospective survey the subject can experience another competing risk such as death (transition to an absorbing state). In such a situation the event history will be recorded only up to the point of death. If death has not occurred the history will be censored at the time of last observation. In this section we will present measures to summarize recurrent event history data without the presence of an absorbing event (death). Modifications to include the competing event in the analysis can be found in Ghosh and Lin (2000).

Suppose that an individual is observed over a time period $[0, \tau_i]$. Let $N(t)$ denote the number of events occurring over the time interval [0, t]. Then the mean cumulative function (MCF) is defined as

$$M(t) = E[N(t)], \text{ where E stands for expectation.}$$

The MCF can also be expressed in terms of renewal density or the occurrence rates. Define renewal density as

$$m(t) = \lim_{\delta t \to 0} \frac{\Pr ob[event\ occurs\ in\ (t, t+\delta t)]}{\delta t}$$

Then $M(t) = \int_0^t m(\tau)d\tau$ or $m(t) = M'(t)$. Following Lawless and Nadeau (1995), we present here nonparametric estimates of mean cumulative function when data on recurrent events are censored.

Suppose an individual i (i = 1,2,.....K) is observed in the time interval $[0, \tau_i)$ and the event times for the individual i are $t_{i1}, t_{i2}, \ldots, t_{ik_i}$. Denote $\delta_i(t) = 1$, if $t \le \tau_i$ and $= 0$, otherwise. Using these notations one can write the total number of individuals at risk to have an event at time t as $r_t = \sum_{i=1}^{K} \delta_i(t)$. Let $n_i(t) \ge 0$ as the number of events for individual i that occurs at time t (usually zero or 1). Total number of events occurring at time t is denoted as $n_{.}(t) = \sum_{i=1}^{K} \delta_i(t) n_i(t)$. Note that if the event times are distinct, $n_{.}(t) = 1$ for all event times. Then the Nelson-Aalen estimator of the mean cumulative function is (Anderson et al., 1993)

$$\hat{M}(t) = \sum_{s=1}^{t} \frac{n_{.}(s)}{r_s} \tag{16}$$

Note that an estimator of occurrence rate can be calculated as

$$\hat{m}(t) = \frac{n_{.}(t)}{r_t} \tag{17}$$

Lawless and Nadeau (1995) provide a robust variance estimator of $\hat{M}(t)$ as

$$\sum_{i=1}^{K}\left\{\sum_{s=1}^{t}\frac{\delta_i(s)}{r_s}[n_i(s)-\hat{m}(s)]\right\}^2 \qquad (18)$$

Assume that event times are distinct (if multiple events occur then we will list uncensored cases first followed by the censored cases). By definition $r_0 = K$ (Note r_0 is the number at risk just prior to the first observed time of occurrence). Then

$$\begin{aligned} r_j &= r_{j-1} \quad \mathit{if} \quad j\ \mathit{is\ an\ event\ time} \\ &= r_{j-1}-1 \quad \mathit{if} \quad j\ \mathit{is\ a\ censored\ time} \end{aligned}$$

With this notation we have

$$MC\hat{F}_0 = \frac{1}{r_0} \text{ and } M\hat{C}F_j = \frac{1}{r_j} + MC\hat{F}_{j-1} \qquad (19)$$

A simple recurrence formula to compute the robust variance can be written as follows:

$$Var_j = Var_{j-1} + \frac{1}{r_j^2}\left[\sum_{i\varepsilon R_j}\left(d_{ij}-\frac{1}{r_i}\right)^2\right] \qquad (20)$$

where d_{ij}=1 if the i^{th} individual had an event at time j, and d_{ij}=0 if the i^{th} individual has no event at time j. Confidence bounds for the cumulative number of events are usually calculated under the assumption that the recurrence time follows a lognormal distribution. Specifically, the lower and upper bound for log (MCF) is calculated as

$$\log(MCF_t)\ \pm z_\alpha \frac{sqrt(Var_t)}{MCF_t} \qquad (21)$$

Exponentiation of the confidence intervals for $\log(MCF_t)$ will give the confidence intervals for MCF_t.

5.1 Examples

In order to illustrate the method we use birth history data of 5 women recorded in a fertility survey at a cross-sectional point.

Table 24.4 Birth history of Woman

Woman	*Ages at birth*
1	15, 17, 18, 20, 25, 30, 32+
2	17, 23, 24+
3	24, 25+
4	23, 28+
5	20, 23+

+ indicates censoring

In Table 24.5 we sort the data by age at birth and censoring. When there is a tie it is assumed that event times precede the censoring time. On the basis of this assumption a risk set is calculated at ages when there is a birth.

Data in Table 24.4 can be sorted in ascending order by age at birth and censoring as shown in Table 24.5. Use equations 17, 19, 20, and 21 to complete the table. Table 24.5 shows that in this sample a woman will have on an average 3.2 children (with a confidence interval of 2.3 and 4.4) by age 30. To get a realistic view, a new dataset with 500 currently married women was created with birth history data similar to Table 24.4. The values of mean cumulative function for selected ages for this data are shown in Table 24.6. The table shows that a married woman in continuous marriage will have, on an average, seven children by age 40 with a confidence interval of (6.5, 7.5).

6 Current status data on recurrent events

Complete data (except for censoring) on recurrent events usually consists of the information of the number of events occurred and the timing

Table 24.5 Calculation of mean cumulative function*

ID	*Age*	*Censoring status*	r_t	$\hat{m}_t$	$\hat{M}(t)$	*Lower limit*	*Upper limit*
1	15	1	5	0.2000	0.2000	0.045719	0.874911
1	17	1	5	0.2000	0.4000	0.140881	1.135713
2	17	1	5	0.2000	0.6000	0.255922	1.406678
1	18	1	5	0.2000	0.8000	0.382493	1.673235
1	20	1	5	0.2000	1.0000	0.516851	1.934792
5	20	1	5	0.2000	1.2000	0.656933	2.192004
2	23	1	5	0.2000	1.4000	0.801453	2.445560
4	23	1	5	0.2000	1.6000	0.949545	2.696028
5	23	0	4				
3	24	1	4	0.2500	1.8500	1.132471	3.022153
2	24	0	3				
1	25	1	3	0.3333	2.1833	1.372788	3.472350
3	25	0	2				
4	28	0	1				
1	30	1	1	1.0000	3.1833	2.315633	4.376081
1	32	0	0				

*For calculation of confidence interval $z_\alpha = 1.65$

Table 24.6 Cumulative mean function (births) for a hypothetical cohort of married women without marital disruption (illustrative data from a demographic survey)

Age	*Risk set*	*M(t)*	*Lower limit*	*Upper limit*
15	500	0.176	0.142844	0.216851
20	459	1.385	1.283716	1.493680
25	306	2.864	2.706247	3.032014
30	159	4.269	4.044511	4.506071
35	61	5.804	5.466983	6.163056
40	20	6.978	6.492089	7.501766

of each event at the time of recording. However, in many occasions the recorded data will consist only of the number of events and the time of recording. For example, some demographic surveys will record only the number of children ever born and the age at survey. Table 24.7 gives illustrative data on number of children born to 47 women, married after age 20 and in intact marriage at the time of survey. The goal of the analysis is to compute the mean cumulative function with the current status data.

Suppose there are k independent subjects and the subjects report $N_i(t)$ events at the time of survey. The goal is to estimate mean cumulative function $M(t) = E(N(t))$ for t = 1, 2, . . . , T, based on the current status data. Suppose that K subjects have different observed times (we will relax this assumption later). Direct information about $M(t)$ is available at the observation times $t_1, t_2, \ldots., t_K$. Note that $M(t)$ should be a non-decreasing function of t. The directly observed $M(t)$ may not be meeting this condition. Sun and Kalbfleisch (1993)

Table 24.7 Distribution of married women by age at survey and number of children ever born

Age at survey	*Distribution of married women by number of children at survey*							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*Total women*
21	1							1
22	3							3
23	5	1						6
24	2	2						4
25	2	3						5
26	2	2	3					7
27	0	1						1
28	0	2						2
29	1	2	2					5
30	0	2	1					3
31	0	0	1					1
32	0	0	1	1	1			3
33	0	1	2	0	0	0	1	4
34	0	0	0	0	1			1
35					1			1

proposed isotonic regression to obtain estimates of $M(t)$ that satisfies the required condition. Specifically, they advocated the pool adjacent algorithm (Barlow et al., 1972) to obtain $M(t)$. The procedure can be briefly described as follows. Rank the observation times and compute the mean response at each time point. If any two adjacent means are out of order $[\hat{M}(t+1) < \hat{M}(t)]$, then the observations in these two means are combined to form a block, and pooled block mean is computed. If any two block means are out of order, the observations in these blocks are pooled to form a new block mean. This process is continued until all block means are in proper order. Confidence intervals for the estimated non-decreasing mean cumulative function can be computed as

$$\max_{i\le l}\{\hat{M}(t_i) - s_K\hat{\sigma}/\sqrt{n_i} \le M(t_i)\}$$
$$\le \min_{j\le i}\{\hat{M}(t) - s_K\hat{\sigma}/\sqrt{n_i}\}$$

where S_k denotes the upper α point of the studentized maximum modulus distribution with parameter k, n_i is the observed number of subjects at time t_i, and $\hat{\sigma}^2$ is the pooled estimate of the variance (Korn, 1982). These confidence intervals are directly obtained from the sample means. Table 24.8 illustrates the method using the data in Table 24.7. Published values of s_k are available in Hahn and Hendrickson (1971). (For large values of k, s_k can be approximated by standard normal distribution.) Observed mean number children at various ages, based on the data, are given in the second column in Table 24.8. Because the observed means are not monotonic, we obtained the pooled estimates shown in column 4 of Table 24.8. For the data the pooled standard deviation is estimated to be 1.318. There were 15 ages represented in the data. For a confidence level of 90% the student's t value is 1.753. The calculated confidence intervals are shown in Table 24.8. The results show that a 35-year-old woman will, on an average, have 5 children with a confidence interval (3.97, 6.03).

Table 24.8 Mean cumulative function and confidence intervals

Current age	*Observed* $\hat{M}(t)$	*Number of women*	*Smoothed* $\hat{M}(t)$	*Lower confidence interval*	*Upper confidence interval*
21	1.0000	1	1.000	0.000	3.310
22	1.0000	3	1.000	0.000	3.310
23	1.1666	6	1.167	0.000	3.310
24	1.5000	4	1.500	0.000	3.386
25	1.6000	5	1.600	0.000	3.386
26	2.1429	7	2.100	0.367	3.721
27	2.0000	1	2.100	0.367	3.721
28	2.0000	2	2.100	0.367	3.721
29	2.2000	5	2.200	0.643	3.757
30	2.3333	3	2.333	0.821	3.846
31	3.0000	1	3.000	1.666	4.334
32	4.0000	3	3.857	2.845	5.155
33	3.7500	4	3.857	2.845	5.155
34	5.0000	1	5.000	3.967	6.033
35	5.0000	1	5.000	3.967	6.033

7 Backward recurrent times

Event history data sometimes record at the time of a survey only the time elapsed since the last event. For example, a fertility survey records the time elapsed between survey and the last live birth. Such data have the advantage that the respondent can accurately recall the timing of the most recent event. Such data is referred to as backward recurrence time data (Allison, 1985). One natural question is whether information on backward recurrence time can be used to summarize the distribution of the interevent times. The standard survival analysis techniques are not directly applicable to this data because, in theory, all observations are censored. Allison (1985) proposed several methods to conduct regression analysis of backward recurrence times. Other applications can be found in Ali et al. (2001), Keilding et al. (2002).

In this paper we look at a nonparametric method to estimate the distribution of the interevent times. Denote $g(y)$ as the density function of the backward recurrence time and $f(x)$ as the density of the corresponding interevent time. When the interevent times have a common distribution with density $f(x)$ (with the corresponding distribution function $F(x)$) relationships between $g(y)$ and $f(x)$ can be established. If the survey time is relatively far from the beginning of the process, the limiting distribution of the backward recurrence time is related to the distribution of the interevent times as:

$$g(y) = \frac{1-F(y)}{\mu}, \text{ where } \mu = E(X), \text{ the mean of the interevent time.}$$

This relationship implies that $\mu = \dfrac{1}{g(0)}$. Using these relationships, a nonparametric estimation of the survival function $S(y) = 1 - F(y)$ is obtained as:

$$\hat{S}(y) = \frac{\hat{g}(y)}{\hat{g}(0)} \qquad (22)$$

Table 24.9 Distribution of backward recurrence time

Time since last live birth (in completed years)	*Proportion of women ($N = 5$)*	*Smoothed proportion $\hat{g}(y)$*	$\hat{S}(y) = \dfrac{\hat{g}(y)}{\hat{g}(0)}$
0	0.110	0.3622	1
1	0.294	0.3008	0.83048
2	0.256	0.2394	0.66096
3	0.168	0.1780	0.49144
4	0.114	0.1166	0.32192
5	0.058	0.0552	0.15240

However, it is shown that the nonparametric maximum likelihood estimator of a decreasing density is inconsistent at 0 (Sun and Woodroofe, 1996). One solution is to estimate $g(0)$ using the observed histogram after smoothing. Usually a smoothing algorithm such as loess (Cleveland and Devlin, 1988) can be used.

7.1 Example

We use some data extracted from a demographic survey that recorded the birth history of women. Five hundred records of women in intact marriage and with at least one birth were extracted. For these women, the backward recurrence (time elapsed (in years) since last live birth at the time of the survey) is noted. The distribution of the observed forward recurrence time is given in Table 24.9. Note that the observed $\hat{g}(0)$ is smaller than the rest of the proportion. Therefore, a smoothed value is calculated. In this case a prediction based on a linear predictor using the time values greater than one is used. The estimate of $\hat{g}(0) = .3622$ is obtained giving a mean birth interval of 2.76 years. The calculated survival function based on equation 21 is presented in Table 24.8. Based on the survival function, the median birth interval is calculated to be 2.95 years.

8 Summary

This paper summarizes a number of nonparametric techniques to describe event history data. No attempt is made to conduct subgroup or covariate analysis. Although summary measures are rather easy to obtain, computation of standard errors may in some situations not be very easy. For example, computations of standard errors of summary measures for multistate data are not very easily obtained. Computation of summary measures and their standard errors are not available in standard statistical packages beyond single spell data without competing risks.

References

Ali, M. M., Marshall, T. and Babiker, A. G. (2001). Analysis of incomplete durations with applications to contraceptive use. *Journal of the Royal Statistical Society*, Series A, 164: 549–563.

Allison, P. D. (1985). Survival analysis of backward recurrence times. *Journal of the American Statistical Association*, 80: 315–322.

Anderson, P. K., Borgan, O., Gill, R. D. and Keilding, N. (1993). *Statistical Methods Based on Counting Processes*. New York: Springer.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions*. New York: Wiley.

Breslow, N. E. (1972). Contributions to the discussions of the paper by D. R. Cox. *Journal of the Royal Statistical Society*, Series B 34: 216–217.

Chiang, C. L. (1968). *Introduction to Stochastic Processes in Biostatistics*. Huntington, NY: Kreiger.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression by local fitting. *Journal of the American Statistical Association*, 83: 596–610.

Ghosh, D and Lin, D. Y. (2000). Nonparametric analysis of recurrent events and death. *Biometrics*, 56: 554–562.

Guo, G. (1993). Event history analysis of left truncated data. In P. Marsden (ed.), *Sociological Methodology,* pp. 217–243. Washington, DC: American Sociological Association.

Hahn, G. J. and Hendrickson, R. W. (1971). A table for percentage points of the distribution of the largest absolute value of k student t variates and its applications. *Biometrika*, 58: 323–332.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation of incomplete observations. *Journal of the American Statistical Association*, 53: 457–481.

Keilding, N., Kvist, K., Hartvig, H., Tvede, M. and Juul, S. (2002). Estimating time to pregnancy from current cross-sectional sample. *Biostatistics*, 3: 565–578.

Korn, E. L. (1982). Confidence bands for isotonic response curves. *Applied Statistics*, 31: 59–63.

Lawless, J. F. and Nadeau, C. (1995). Some robust methods for the analysis of recurrent events. *Technometrics*, 37: 155–168.

Namboodiri, N. K. and Suchindran, C. M. (1987). *Life Table Techniques and Their Applications*. Orlando, Fl: Academic Press.

Sun, J. and Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Association*, 88: 1449–1453.

Sun, J. and Woodroofe, M. (1996). Adaptive smoothing for a penalized NPMLE of a non-increasing density. *Journal of Statistical Planning and Inference*, 52: 143–159.

This page intentionally left blank

Chapter 25

The Cox proportional hazards model, diagnostics, and extensions

Janet M. Box-Steffensmeier and Lyndsey Stanfill

Event history analysis (also referred to as duration, survival, or reliability analysis) is a technique that allows researchers to assess the implicit risk of an event occurring. That is, we consider not only whether an event occurs, but when. Event history analysis provides an understanding of the timing and history leading up to the event, from which we can draw inferences about the process. However, we need to check several diagnostics, in order to have confidence in our inferences. This chapter discusses the Cox proportional hazards model and the necessity for diagnostics. Cox modeling extensions to account for dependence, such as the conditional frailty model, are also presented.

1 Introduction

An event is a change from one state to another; examples include death, marriage, fall of governments, or dissertation completion. The dependent variable in an event history model is the time until the event occurs. Event history models are well-suited for longitudinal analysis because they can easily accommodate problems common to longitudinal data such as censored observations and time-varying covariates (explanatory variables).[1] Censoring occurs when information about an observation is incomplete, such as when an observation has not changed from one state to another when the data collection process ends. Time-varying covariates can take on different values over time for a single observation. Censoring and time-varying covariates often present statistical problems that can be overcome by a survival model.

Two types of event history models are parametric and semiparametric models. Parametric models assume that the time until an event occurs follows a specific distribution, such as the exponential, and the distribution of when the events happen can be thought of as time dependency in the data. Parametric distributions are most often used in engineering when the analyst has a strong understanding of the distribution of the risk of failing with respect to time. The primary advantage of parametric event history models is the ability to forecast.

[1] Klein and Moeschberger (1997), Hosmer and Lemeshow (1999), Therneau and Grambsch (2000), Blossfeld and Rohwer (2001), Singer and Willet (2003), and Box-Steffensmeier and Jones (2004) all provide useful texts on event history.

Conversely, semiparametric models do not specify a distributional shape for the timing of events; rather semiparametric models are parameterized by the explanatory variables. A semiparametric model is more appropriate when the primary objective is to understand the impact of covariates on the risk of an event. In this situation, duration dependence is considered a nuisance. Time-dependency can be thought of as the "left over" effects of time after the hazard rate has been conditioned by the covariates. If the model had been perfectly specified, there would be no time-dependency because the hazard rate would be fully characterized by the covariates.

The Cox proportional hazards model is the most commonly used semiparametric event history model, and this model estimates the impact of covariates on the risk of an event without *a priori* specifying a specific distribution for the duration dependence or making any assumptions about when an event occurs. The inferences drawn from the estimation may be misleading if the distribution of parametric model is incorrectly specified.[2] Rather, the hazard rate is parameterized only by the covariates of interest. Typically, in social science research, theory is not strong enough to correctly specify a distribution shape for the timing process, and thus the less restrictive Cox model is preferred. One advantage of the Cox model is the wide variety of diagnostic tests that have been developed.

2 Cox proportional hazards model

Key concepts for estimating and understanding the Cox model are the hazard rate, risk set, and survival function. A hazard rate can be thought of as the probability that an event will occur for a particular observation at a particular time, or the rate at which an event occurs for an observation at time t given that the observation has survived through time $t-1$. In the Cox model, the hazard rate for the ith individual is given by:

$$h_i(t) = h_0(t)\exp(\beta'\mathbf{x})$$

The baseline hazard rate, $h_0(t)$, is a constant (unspecified) baseline hazard rate, and $\mathbf{x}$ is a vector of covariates. A Cox model does not report an intercept; instead the intercept is absorbed into the baseline hazard function. However, should the researcher have a need for the baseline hazard rate, it can be calculated from the estimates. The underlying hazard rate can be thought as the hazard rate when all of the covariates equal zero. Therefore, any change to the hazard rate is a function of the values of covariates. Hazard rates are substantively interesting to researchers who seek to understand how an event is conditional on its history.

The risk set includes all of the observations that are still at "risk" for experiencing the event. Risk is an implicit aspect of the Cox model because the hazard rate is derived from the risk set. Once an observation experiences the event at time t (changes from one state to another) it drops out of the risk set and is no longer part of the dataset being analyzed in later periods ($t > T$). Instead, the observation is now incorporated into the failure rate. In survival models, the hazard rate is a ratio of failure times and survival times:

$$h(t) = \frac{f(t)}{S(t)}$$

where $h(t)$ is the hazard rate, $f(t)$ is the failure rate, and $S(t)$ is the survival function. When an object is still at risk, it is incorporated into the model in the survival function. This method is what allows for survival models to uniquely integrate censored observations into the model. An observation that is censored is simply still

[2] See Box-Steffensmeier and Jones (2004) and Golub (forthcoming) for a more elaborate discussion of the advantages of the Cox model.

in the survival function at the end of the period of study.

3 Cox model residuals

Unlike least-squares residuals, which are the distance between the observed and predicted values of an observation, a duration model does not always provide a systematic component for censored observations due to estimation via partial maximum likelihood (Box-Steffensmeier and Zorn, 2001; Hosmer and Lemeshow, 1999). However, several different kinds of residuals have been developed for assessing the adequacy of the model. Some of the most common and useful residuals are:

- Cox-Snell residuals used to assess the overall fit of a posited model
- martingale residuals used to assess the functional form of covariates and to compute additional residuals
- score residuals used to assess the potential influence of an observation on the estimated coefficients
- deviance residuals used to detect outliers
- Schoenfeld residuals, which are critical for testing the proportional hazards assumption.

Cox-Snell residuals are based on the expected number of events in a given time interval or an expected count. Recall from above that the Cox model,

$$h_i(t) = h_0(t)\exp(\beta'\mathbf{x})$$

estimates survival times $\hat{S}_i(t)$. If the model is adequate the estimated survival times $\hat{S}_i(t)$ should be very similar to the actual survival times $S_i(t)$. Cox-Snell residuals assess the relationship between the estimated and actual survival times. The residual is given by:

$$r_{cs_i} = \exp(\hat{\beta}'\mathbf{x}_i)\hat{H}_0(t_i)$$

where $\hat{H}_0(t_i)$ is the cumulative hazard. If the correct model has been fit to the data, then r_{cs_i} has a unit exponential distribution. This means that the hazard ratio equals one (for further discussion, see Box-Steffensmeier and Jones, 2004; Collett, 1994; Cox and Snell, 1968; Klein and Moeschberger, 1997). The Cox-Snell residuals are most often used to assess how well the model fits the data, which is discussed later in the chapter.

The Martingale residuals use a "counting process" approach. To understand the intuition, Therneau and Grambsch say to think of each observation "… as the realization of a very slow Poisson process" where "censoring is not incomplete data, rather the Geiger counter just hasn't clicked yet" (2000, p. 68). This concept has created a way for researchers to overcome the problem of not having an actual failed time for every observation.

The counting-process representation of the Cox model is a linear-like model that counts whether the event occurs a time t:

$$\delta_i(t) = H_i(t) + M_i(t)$$

and rearranging terms:

$$M_i(t) = \delta_i(t) - H_i(t)$$

To explain, $\delta_i(t)$ is a censoring indicator. Each observation receives a zero for the censoring indicator until it experiences an event. In the time period the observation experiences the event, $\delta_i(t) = 1$, and equals one for every time period afterwards. $H_i(t)$ is the hazard or the risk of the event occurring for an observation at each time period. After an observation experiences an event, $H_i(t) = 1$ for each time period after the event.

$M_i(t)$ is referred to as the *martingale* and can be thought of as the error component. The martingale has the same properties as the least-squares residuals: the mean of the residuals is zero ($E(\mathrm{Mi}) = 0$) and there is no covariation in the residuals across obsevations ($\mathrm{cov}(M_i, M_j) = 0$). The martingale is *equivalent* to the censoring indicator minus the Cox-Snell residuals

$$M_i(t) = \delta_i(t) - r_{cs_i}$$

Thus, we see that the martingale residuals can be used to compute other types of residuals, but the martingale residual can also be used to assess the functional form of a covariate.

Martingale residuals can also be used to create score residuals for each covariate. Score residuals can then be used to assess the potential influence of an observation on the estimated coefficients. Score residuals for the *i*th subject on the *k*th covariate are calculated as:

$$L_i = \int_0^\infty [X_i(t) - \overline{X}_i(t)]d\hat{M}_i(t)$$

where $\overline{X}(t)$ is the weighted mean of covariate X over the observations still in the risk set at time *t*. The weight, $d\hat{M}_i(t)$, is the change in the martingale residuals for the *i*th subject and time *t*.

Unlike least-squares residuals, martingales are not distributed around zero. Deviance residuals are also martingale based and normalize the martingale residuals so that they are symmetric around zero. These residuals are calculated using using:

$$D_i = sign(M_i(t))\{-2[M_i(t) + \delta \log t(\delta_i - M_i(t))]\}^{1/2}$$

$M_i(t)$is the martingale residual for the *i*th observation, and where the martingale is zero, the deviance residual is zero.

Finally, Schoenfeld residuals are needed to test the proportional hazards assumption. The Schoenfelds are simply the sum across observations of the score residuals for each covariate *k*:

$$S_{kt}(t) = \sum_{i=1}^{N} L_{ki}(t)$$

The Schoenfeld residuals can be thought of as the observed minus the expected values of the covariates at each failure time, and the summation above yields a single value for each covariate at each time point.

$$\hat{r}_{sik} = c_i(x_{ik} - x_{wik})$$

These residuals can be used to assess the proportional hazards assumption.

The Cox model is generally a very robust model, however, as in any model, diagnostic checking is important. The residuals are used for a variety of these diagnostics, including linearity in the covariates and proportional hazards.

4 Covariate functional form

As with least-squares models, the functional form of the covariates must be tested. The Cox model assumes that the covariates are loglinear, but often continuous variables assume a more complicated form. Given the nature of event data, nonlinear functional forms are more likely in event history models than standard least-squares models.

Failing to detect and correct for nonlinearity leads to several undesired effects in the model (Therneau and Grambsch, 2000; Keele, 2005). First, the estimates are biased and exhibit decreased power of statistical tests. Second, the fact that the effect of the covariate differs across the values of the explanatory variable changes the interpretation of the impact of the variable on the hazard rate. Finally, failing to detect nonlinearity has consequences for diagnosing and correcting violations of the proportional hazards assumption. Specifically, tests for nonproportionality can fail in the presence of an incorrect functional form, and correcting for a violation of the proportional hazards assumption when the failure is due to nonlinearity will not produce the correct model. Thus, the functional form of the covariates should be assessed prior to testing for nonproportionality (Therneau and Grambsch, 2000; Keele, 2005). Three methods are available for detecting nonlinearity in the covariates. The first plots the martingales saved from a Cox model against each covariate. An alternative test uses a two-step process in which martingales are saved from a m-1 Cox model (where m is the number of covariates in the model). Then those smoothed martingales would be plotted against the missing covariate. The process is repeated for each covariate. The third method for assessing the linearity of covariates directly models the functional form with a smoothing spline and provides a statistical test for the presence of nonlinearity (Keele, 2005). The model is estimated using smoothing splines for any covariates suspected of nonlinearity (either because of theory or because of diagnosis using martingale plots). Then, a Wald test can be used to decide whether a nonlinear affect should remain in the model.[3]

We illustrate the importance of assessing covariate functional form of the Cox model using data on international disputes among 827 “politically relevant” dyads from the period of 1950 to 1985 (Oneal and Russett, 1997; Beck, Katz and Tucker, 1998; Box-Steffensmeier and Zorn, 2001). Each dyad (pairs of nations) is observed once for each year it is in the dataset for a total of 20,448 observations. The dependent variable is the duration from the beginning of the period until the onset of a militarized international dispute between the two nations that make up the dyad. For pedagogical reasons, we model duration as function of just three factors: the level of democracy in the dyad (scaled from most autocratic (0) to most democratic (1)), the presence of an alliance between the nations (binary variable where 1 indicates an alliance), and whether the two nations are geographically contiguous (binary variable where 1 indicates contiguity).[4] The model has only one continuous variable to test for nonlinearity. After estimating the Cox model, we plot the saved martingales against the democracy covariate using a lowess smoother. Upon examining the plot (Figure 25.1), it is clear that this covariate violates the linearity assumption as it is not a straight line. The alternative two-step test confirms the violation (Figure 25.2). Based on the figures, a quadratic transformation of the democracy covariate is suggested. Now that the linearity of the covariates has been tested, we are able to perform tests of the proportional hazards assumption.

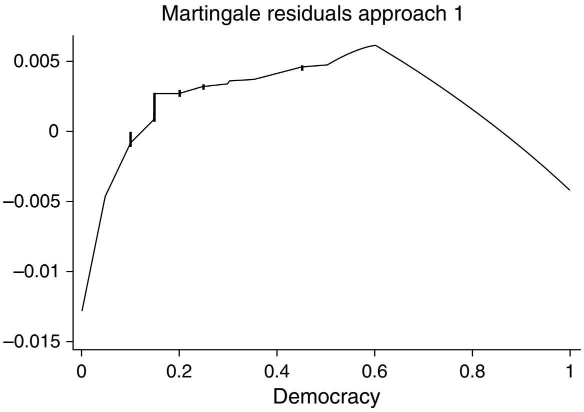


Figure 25.1 Test of democracy covariate functional form

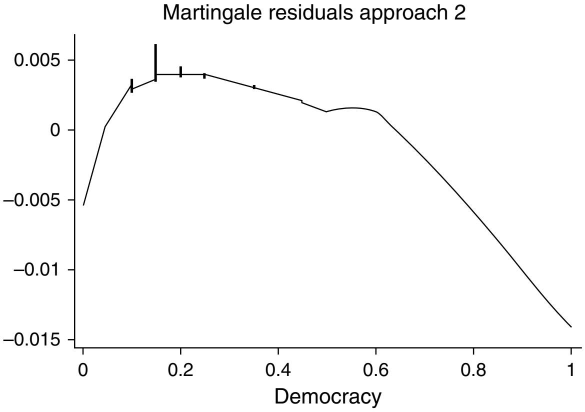


Figure 25.2 Two-step test of democracy covariate functional form

[3] See Keele (2005) for an in depth discussion of directly modeling functional forms using smoothing splines in R.

[4] The full model can be found in Box-Steffensmeier and Jones (2001) or Box-Steffensmeier and Zorn (2001).

5 Proportional hazards assumption

The Cox model assumes that the hazards of two observations with different values on one or more covariates differ only by proportionality (Box-Steffensmeier and Zorn, 2001). If the proportional hazards assumption holds, the hazard value will not differ as a function of time: a hazard rate will be the same at the first time period under study as it is in the last period under study.

To illustrate proportional hazards, suppose we have two observations, A and B, in our data. The hazard rates at time t for observation A, $h_A(t)$, and observation B, $h_B(t)$, are proportional for any value of t. This can be expressed as:

$$h_A(t) = Ch_B(t)$$

where C is a nonnegative constant equal to the proportion of the two hazards, which can also be shown as

$$C = \frac{h_A(t)}{h_B(t)}$$

This assumption implies that the ratio of two hazards is constant over time. The effect of a covariate shifts the hazard rate by a factor of proportionality regardless of when the event occurs (Box-Steffensmeier and Zorn, 2001).

As noted above, the hazard rate for the Cox model is given by:

$$h_i(t) = h_0(t)\exp(\beta'\mathbf{x})$$

Since the hazard rate for the Cox model is proportional, the ratio of two hazards can be written as:

$$\frac{h_i(t)}{h_0(t)} = \exp[\beta'(x_i - x_j)]$$

The proportional difference between the two observations is a function of having different values for the covariates.

Assessing whether or not the proportional hazards assumption holds is essential when estimating a Cox model. Violation of the assumption can lead to biased estimates and decreased power in statistical tests (Box-Steffensmeier and Jones, 2004). A hazard rate that is increasing over time tends to overestimate the impact of covariates. Alternatively, a hazard rate that is decreasing over time, or converging, is biased towards zero (Kalbfleisch and Prentice, 1980).

There are numerous substantive reasons we may not expect the assumption of proportional effects to hold. We may hypothesize that the effect of one of the covariates changes over time due to factors such as learning, life-course changes, or institutionalization. For example, life-course changes may lead us to posit that the effect of unemployment on recidivism varies over time. Or the process of institutionalization may lead us to expect that political alliance size may be large early in the duration of an alliance, but decrease over time (Zorn, 2000).

The generalized Cox model, which allows hazard ratios to vary over time, can be expressed as:

$$h(t) = h_0(t)\exp[X_i\beta + (X_i g(t))\gamma]$$

where the effects of individual covariates are allowed to vary by some function $g(\cdot)$ of time. Tests of proportional hazards assume that $\gamma = 0$, or that any change in the hazard rate is a function only of the covariates.

Violations of the proportional hazards assumption are detected with residual-based tests. The first test for proportionality uses the Schoenfeld residuals. If proportional hazards holds, there should be no relationship between an observation's residual for that covariate and the length of the survival time. A plot of the Schoenfeld residuals against time reveals whether the values of $\hat{r}_{sik}$ are changing with respect to time.

Returning to the interstate dispute example from above, we estimate a Cox model and save the Schoenfeld residuals. Of the three covariates, only the allies plot looks as though the covariate might exhibit nonproportionality due to the plotted line not being straight (see Figure 25.3). However, plots can be misleading and lack a clear diagnosis of nonproportionality; therefore, statistical tests are recommended in addition to residual plots.

Terry Therneau, Patricia Grambsch and Thomas Fleming (1990) first developed a statistical test based on Schoenfeld residuals to detect a global violation of the proportional hazards assumption, i.e., a violation of the proportional hazards assumption for the model as a whole. This global test uses the maximum of the absolute value of the residuals summed over time. A second residual-based statistical test evaluates the proportionality of each covariate; Harrell's (1986) ρ is the correlation of each covariate's Schoenfeld residuals and rank of survival time. Statistical significance is based on the chi-square distribution where the null is $\rho = 0$ or no correlation between the residuals and time. Grambsch and Therneau (1994) modify this test by using the scaled residuals, and also provide an improved global test for nonproportionality based on the aggregated (across covariates) covariance between the unscaled Schoenfeld residuals and survival time.

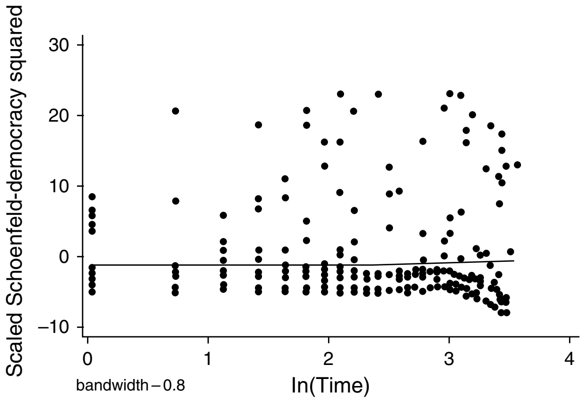


Figure 25.3 Test of PH assumption for democracy

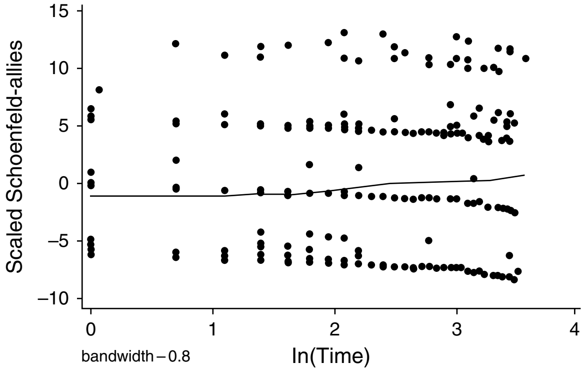


Figure 25.4 Test of PH assumption for allies

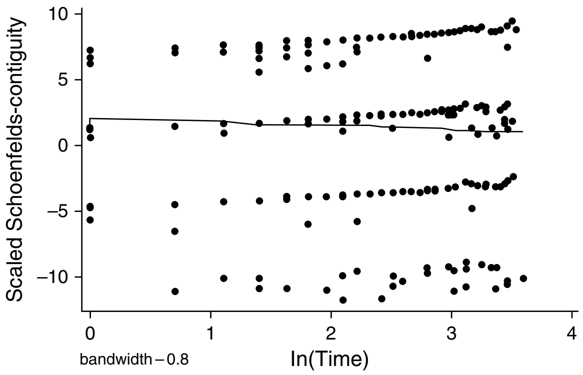


Figure 25.5 Test of PH assumption for contiguity

Table 25.1 presents the statistical tests of the proportional hazards assumption for the interstate disputes example. The columns designated ρ report the estimated correlation between the scaled residuals and *ln*(*Time*),

Table 25.1 Statistical tests of the proportional hazards assumption

Covariates	ρ	χ^2	*d.f.*	*p-value*
Democracy	0.086	4.36	1	0.0368
Allies	0.146	23.47	1	0.0000
Contiguity	−0.052	3.04	1	0.0813
Global test		26.19	3	0.0000

Table 25.2 Comparing Cox regression estimates

	Original model	*Model correcting for nonlinearity*	*Model correcting for nonlinearity and nonproportionality*
Covariates	*Coefficient (s.e.)*	*Coefficient (s.e.)*	*Coefficient (s.e.)*
Democracy	−1.03 (0.22)*		
Democracy^2		−1.20 (.24)*	−1.62 (.45)*
Democ^2 × ln(Time)			.17 (.17)
Allies	−0.37 (0.17)*	−0.37 (.17)*	−1.11 (.29)*
Allies × ln(Time)			.35 (.11)*
Contiguity	1.47 (0.17)*	1.48 (.17)*	1.48 (.17)*
R^{20}	0.32	0.33	0.30
lnL	−2523.09	−2517.93	−2512.20
N	20448	20448	20448

Note: Efron method used for ties. Coefficients are Cox proportional hazards estimates with robust standard errors in parentheses. One asterisk indicates $p < .05$.

while the χ^2 and *p*-values indicate the confidence with which we can reject the null hypothesis that the hazard ratios for different values of that covariate are constant over time. The global test shows a problem with the proportional hazards assumption, ($p < .001$). In addition, quadratic democracy covariate and the allies covariate both have *p*-values lower than conventionally accepted levels ($p < .05$). In this example, the residual plots might have been misleading because the plot of democracy on time did not appear nonproportional.

To correct for nonproportionality, the offending covariate is interacted with some function of time; usually the interaction is ln(Time). Both the offending covariate and the covariate interacted with time are included in the new model.[5] Both the quadratic democracy covariate and the allies covariates are interacted with ln(Time) and included in the new Cox model. Table 25.2 presents the Cox model with nonproportionality and the Cox model with log-time interactions.

When we examine the original model against models that correct for nonlinearities and nonproportionality, it is easy to see that failing to test the assumptions of the model could have serious consequences. While the variables retained the same direction and statistical significance, the impact of the covariates on duration times does change (see Table 25.2). The coefficient for democracy changes from −1.03 to −1.20 when accounting for the correct functional form and to −1.62 when accounting for functional form and nonproportionality. Similarly, the coefficient for allies changes from −0.37 to −1.11 when accounting for nonproportionality. In addition to testing linearity and proportional hazards assumptions, the researcher should perform diagnostics for outliers and leverage and can assess the fit of the model.

6 Other diagnostics

Residual-based diagnostics have been developed to test for outliers, influence, and the

[5] Testing the statistical significance of time-interacted terms has been posited as a third diagnostic technique for detecting nonproportionality. However, using this method to detect and correct for violations of the assumption is not recommended as it is the approach for correcting the problem as well (Box-Steffensmeier and Jones, 2004; Grambsch and Therneau, 1994).

adequacy of the model. We can use the deviance residuals to assess the model in terms of the *i*th observation to identify outliers. Outliers are problematic for the model because they can lead to erroneous conclusions about the hazard rate. A plot of deviance residuals against the observation numbers demonstrates which observations need to be examined more closely. In addition, a plot of deviance residuals against duration time can provide some initial insight into the adequacy of the specified Cox model (see Box-Steffensmeier and Jones, 2004, for elaboration).

Score residuals, on the other hand, are used to measure the influence of an observation on the size of the coefficient. An observation with influence on the size of the coefficient tempers the claims that a researcher can make from the model. The matrix of score residuals, along with the variance–covariance matrix, creates a measure analogous to the dfbeta used with least-squares models. Dfbetas measures the influence of the *i*th individual on the *j*th covariate. In other words, multiplying the score residuals by the variance–covariance matrix provides a measure for how much each observation increases or decreases a given covariate.

Cox-Snell residuals are used to assess the fit of the model. Recall from above that Cox-Snell residuals can be thought of as the expected number of events in a given time interval or the expected count. Therefore, in using these residuals we can better understand how well the model matches the data. Since the Cox-Snell residuals, r_{csi} , follow a unit exponential distribution, a plot of the residuals on the integrated (or cumulative) hazard rate based on the residuals should yield a straight line through the origin with a slope equal to 1 (a 45-degree angle). In the example of interstate dispute, the plot of the Cox-Snell residuals against the cumulative hazard rate of the residuals could be considered a concerning lack of fit (see Figure 25.6). We should be concerned that the model has not been well specified.

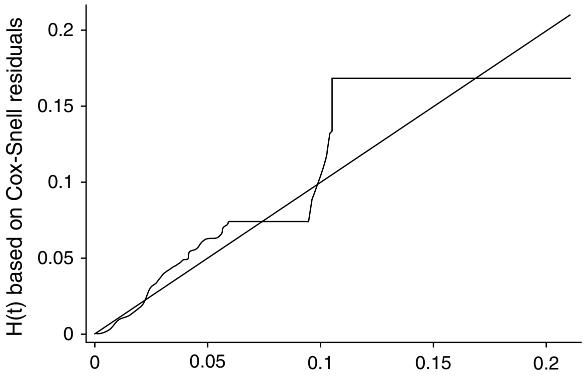


Figure 25.6 Cox-snell residuals from interstate dispute data

While the plot of the Cox-Snell residuals is commonly used for evaluating the fit of the model, alternative techniques are being developed to indicate how well the model performs. A new R^2 statistic for survival models allows us to assess the goodness of fit by measuring the amount of explained variation. In duration data, researchers are interested in how much of the variation in the survival time is accounted for by the model. However, unlike least-squares models which do not have censoring, the R^2 statistic for Cox models must take into account the number of uncensored observations (observations experiencing the event) (Royston, 2006).[6] However, a higher R^2 does not necessarily indicate that a model better fits the data. Rather, the R^2 provides an understanding of how the model accounts for the variation in survival times. In the interstate dispute example, $R^2 = 0.30$, which we interpret as 30% of the variation in survival time, can be explained by the model. When used with other measures of model fit, researchers can use this statistic for a better understanding of the adequacy of the model.

[6] See Royston (2006) for an in-depth explanation of the computation of the R^2 for survival models. Royston provides syntax for using the statistic in STATA.

7 Interpreting a Cox model

The Cox model coefficients are parameterized in terms of the hazard rate $h_i(t) = h_0(t)\ \exp(\beta' \mathrm{x})$, where $h_0(t)$ is the baseline hazard rate and $\beta'\mathrm{x}$ are the covariates and regression parameters. Thus, a negative coefficient indicates that the hazard is decreasing as a function of the covariate and survival time is increasing. Conversely, a positive coefficient indicates that the hazard is increasing as a function of the covariate and survival time is decreasing.

In the example of interstate disputes, the coefficients for democracy and allies are both negative, indicating that the hazard is decreasing as the values of the covariates are increasing, and survival times are increasing. Conversely, a positive coefficient for contiguity indicates that the hazard is increasing when the dyad is contiguous.

Since the parameter estimates for the Cox model reveal information regarding the hazard rate, we can provide a more substantive interpretation of the findings. First, the hazard ratio can be found by exponentiating the hazard rate. A hazard ratio of less than 1 indicates that risk decreases as the covariate increases, and a hazard ratio of greater than 1 indicates that the risk increases as the covariate increases. Further, a hazard ratio close to one suggests that the hazard rate does not change as a function of the covariate. Hazard ratios are readily interpretable for binary covariates. For example, the allies covariate of −1.11 indicates that the hazard is decreasing as a function of the covariate and that survival times are increasing. Or, interpreted as a hazard ratio, the risk of an interstate dispute when the dyads are allies is .33 (exp(−1.11)) lower than when dyads are not allies. However, when states are contiguous, the risk of interstate dispute is 4.39 (exp(1.48)) times greater than when the nations are not contiguous.

An additional tool for interpreting the Cox model examines the percent change in the hazard rate as a function of the covariate. The percent change is calculated as:

$$\%\Delta h(t) = \left(\frac{\exp(\beta(x_i = X_1) - \exp(\beta(x_i = X_2)}{\exp(\beta(x_i = X_2)}\right) * 100$$

where x_i is the covariate and X_1 and X_2 are different values of the covariate. For example, contiguity with a coefficient of 1.48, as above, impacts the hazard rate with a increase of 339%:

$$\%\Delta h(t) = \left(\frac{\exp(1.48(1)) - \exp(1.48(0))}{\exp(1.48(0))}\right) * 100 = 339\%$$

When examining the substantive impact of the other covariates, however, we need to take into account the time-interactions included for nonproportionality. When the coefficients for a covariate and a time-interacted covariate have the same sign, the hazard ratios diverge over time, but when the signs are opposite, the hazard rates converge and then diverge (Teachman and Hayward, 1993). The percent change in hazard rate for time-interacted covariates can be calculated from:

$$\%\Delta h(t) = \left(\frac{\begin{array}{l}\exp[(\beta_{k1}(x_i = X_1) + \beta_{k2}(x_i = X_1)\ln(T)] \\ -\exp[(\beta_{k}(x_i = X_2) + \beta_{k2}(x_i = X_2)\ln(T)]\end{array}}{\exp[(\beta_{k}(x_i = X_2) + \beta_{k_2}(x_i = X_2)\ln(T)]}\right) * 100$$

where β_{k1} is the coefficient for the original covariate, β_{k2} is the covariate for the interaction, and T is a specific time. This calculation takes into account the change of the hazard rate over time.

For example, the allies covariate was interacted with the log of time due to violations of the proportional hazards assumption. Substantively, the impact on the hazard rate of a dyad being an ally would be a 7% decrease without taking into account the time interaction. When both allies and ln(t) allies are considered, the

Table 25.3 Cox regression of the timing of interstate disputes

Covariates	*Coefficient (s.e.)*	*Hazard ratio*	*Change in X*	*Impact of the covariate on the hazard rate*
Democracy^2	−1.62*(.45)	0.20	0,1	−77% (T = 1)
Democ^2 × ln(Time)	.17(.17)	1.19		8% (T = 10)
Allies	−1.11*(.29)	0.33	0,1	−53% (T = 1)
Allies × ln(Time)	.35*(.11)	1.42		90% (T = 5)
Contiguity	1.48*(.17)	4.39	0,1	339%
R^2	0.30			
lnL	−2512.20			
N	20448			

Note: Efron method used for ties. Coefficients are Cox proportional hazards estimates with robust standard errors in parentheses. One asterisk indicates $p < .05$.

percent change in the hazard rate from a dyad that is allied compared to a dyad that is not allied is decreased 53% at T = 1 but increases 90% at T = 5. Table 25.1 provides a substantive explanation of the impact of each covariate on the hazard rate.

A final way of substantively interpreting the model is to examine the baseline hazard rate. The baseline hazard rate is not calculated directly by the Cox model, but can be retrieved in the event that the researcher is interested. Graphing the baseline hazard rate can provide a useful illustration of the model (see Box-Steffensmeier and Jones, 2004, for further explanation); however, researchers are usually more interested in the impact of the covariates on the hazard. An important substantive motivation for using duration models is an implicit interest in risk or timing of an event. Providing a substantive interpretation of the model results provides the researchers with the tools to fully characterize the process under examination.

8 Cox modeling extensions and sources of dependence

Thus far, we have discussed the Cox model with respect to single events; however, the flexible Cox model is amenable to multiple events data. The most important categories of multiple events to consider are unordered and ordered. Unordered multiple events refer to the situation where important substantive distinctions are drawn about the event. For example, it is not just that a cabinet government failed but how it failed, e.g., it was dissolved and early parliamentary elections were called or the incumbent cabinet is directly replaced by a new one (Diermeier and Stevenson, 1999). Or for a study of unemployment duration, Addison and Portugal (2003) argue that it is important to distinguish exit from unemployment by finding a job or inactivity. When unordered multiple events are considered, the Cox model is stratified by event type. This allows the baseline hazard to differ by event type. Each stratum has its own baseline hazard function, while the covariates are constrained to be the same across the different strata. This model is the well-known competing risks model.

Ordered multiple events are generally referred to as repeated events. For example, patients may suffer multiple heart attacks, criminals may return to prison numerous times, or countries experience multiple civil wars. Event history models for repeated events explicitly incorporate the reality that the risk of experiencing an event may change once the event has already been experienced. That is, previous event occurrence impacts the hazard of experiencing the event again, so analysts cannot assume the events are independent.

In repeated events data, two types of dependence are possible: event dependence and case dependence. Event dependence occurs when an event is conditional on and influenced by a previous occurrence. Experiencing one heart attack may weaken the heart and make the individual more likely to suffer another. Detecting event dependence in the model is important because the risk of later events may be conditioned by a previous experience, thus a hazard rate that does not stratify does not allow for the influence of an earlier event to change the hazards for later events. Case dependence, also referred to as heterogeneity, occurs because observations in repeated processes are correlated. There may be unmeasured, unmeasurable, or unimagined factors that affect whether or not the observation experiences the event. For example, it is widely held that culture and history affects the likelihood of a country experiencing civil wars. This heterogeneity is important to take into account for accurate inferences. If we do not account for heterogeneity in the model, i.e., if we treat all observations as independent, we overstate the amount of information provided by each case and produce incorrect standard errors. When analyzing repeated events data, we want a method that can detect and correct for event dependence and heterogeneity.

The conditional frailty model is equipped to handle both event dependence and heterogeneity (Box-Steffensmeier and DeBoef, 2006). The model accounts for event dependence by stratifying and for heterogeneity by incorporating a frailty term. The model is presented as:

$$\lambda_{ik}(t) = \lambda_{0k}(t)\exp(X_i(t)\beta + \mu_i)$$

where $\lambda_{ik}(t)$ is the case's risk for event k as a function of an event specific baseline hazard, $\lambda_{0k}(t)$, and a case specific random effect, μ_i. If the model exhibits event dependence, the event specific baseline hazards will show separation. If the model exhibits heterogeneity, the frailty term, or random effect, is statistically significant. Separately incorporating event dependence and heterogeneity into the model allows the sources of dependence to be separated and correct inferences to be made about the effect of the covariates in the model.

We illustrate the conditional frailty model using data on the duration spent in foster care. If heterogeneity is the underlying problem, then efforts to reduce event rates are best spent searching for ways to target the needs of specific types of children with disproportionate use of the system and to change the conditions associated with churning, i.e., placement instability. Churning is an important policy consideration because it has been linked to weakened attachment to a child's primary care giver and to emotional and behavioral problems as well as school failure, criminal activity, and early parenthood (Cook et al., 1991; Fanshel et al., 1990; Goldstein et al., 1973; Leiberman, 1987; Zimmerman, 1982). If event dependence is itself quite high, then just being in the system fosters further time in the system and legislation designed to limit time and multiple placements, such as the Adoption and Safe Families Act of 1997, should be quite effective. In addition to disentangling the effects of event dependence and heterogeneity, the model is critical for obtaining accurate assessments of the effects of measured covariates like sex or age. The model is underspecified, but useful for pedagogical purposes to illustrate not only why we may be interested in separating out event dependence and heterogeneity, but how to interpret such results.

We use foster care data from the State of Tennessee's Department of Children's Services, obtained with the help of the Chapin Hall Center for Children at the University of Chicago. We currently have data on children placed for the first time in 2000 and 2001. Their placement histories are observed through December 31,

2003. Table 25.4 shows that the variance of the random effect is statistically significant, providing evidence of heterogeneity across individuals. Controlling for heterogeneity allows us to make correct inferences about covariates in the model. Evidence of event dependence is also apparent. Figure 25.7 shows the cumulative baseline hazards, which vary by event number even after accounting for heterogeneity via a random effect. This suggests that placements are event dependent, with more frequent placements leading to further disruptions and more movement via foster care placements.

Table 25.4 Conditional frailty model, placements

Covariate	*Estimate*	*Robust SE*
urban	−0.209	0.034
gender	0.061	0.020
black	0.168	0.025
hispanic	0.012	0.054
other	−0.075	0.051

Variance of random effect = 0.15, statistically significant p-value of 0.00

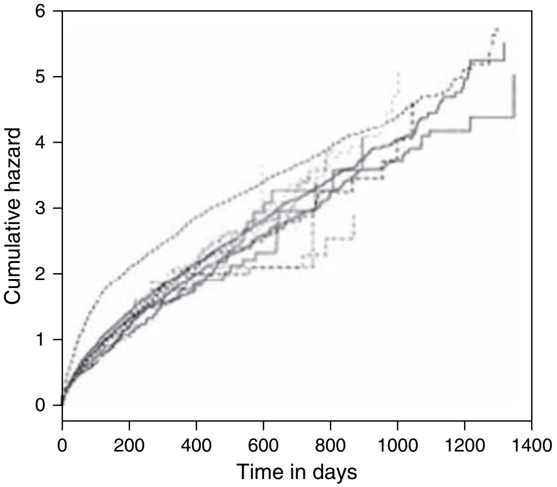


Figure 25.7 Conditional (gamma) frailty model: placements data

Another useful extension of the Cox model is a multilevel approach. Both a basic Cox model and the conditional frailty model can be a multilevel model. For example, we could examine differences across the twelve administrative regions in Tennessee or add data for additional states and examine state policy differences. For example, Gifford and Foster (2005) find in their cross-classified, multilevel model, that facility-level factors are a key determinant of inpatient length of hospitalization. Facility-level factors explain a greater proportion of the overall variation than do even individual characteristics (Gifford and Foster, 2005). So, appropriate analyses may need to allow for the fact that duration spells can be nested within individuals and perhaps facilities/providers, schools, congressional districts, states, etc.

Finally, spatial dependence is another type of dependence that may arise in duration models and is a particularly promising extension. Subjects may be dependent upon observations related in spatial proximity. For example, neighboring states may influence one another or be dependent on one another. Standard survival models estimate the impact of variables on the timing or risk of an event but do not, however, provide a rigorous or generalized mechanism for modeling this spatial dependence. In the past, survival models have incorporated a spatial element via a dummy variable for contiguity or a proportional measure of the number of neighbors previously experiencing the event (Berry and Berry, 1990; Volden, 2006). While these methods attempt to theoretically incorporate spatial dependence, they do not capture the simultaneity and multidirectionality of spatial dependence. Moreover, the lagged proportional measure and spatial influence is conditional and unidirectional, and using this type of measure when the process is simultaneous and multidirectional results in biased estimates of spatial influences (Anselin, 1988). Bayesian spatial survival models incorporate spatial dependence via a frailty term. In these models, the unobserved shared risk of experiencing an event is parameterized as a function of spatial proximity between neighboring observations. Darmofal and Young (2006) provide innovative work assessing the adequacy of duration models in the face of spatial dependence.

9 Conclusion

In using the Cox model, researchers have opened a window to investigating the process of events. By understanding the timing of events and how covariates impact timing, we better understand how the process unfolds. Traditional models may indicate what variables are statistically significant for the occurrence of an event, but some of those variables may cause an event to occur much faster than others. The Cox model is a flexible and robust tool that equips the researcher for investigating these processes. Performing rigorous diagnostics ensures that the model produces accurate inferences into the process at hand. Furthermore, state-of-the-art extensions of the Cox model, such as the conditional frailty model and the Cox model with spatial dependence, open the field even wider. Armed with these new techniques, researchers provide insights into more complicated event processes.

References

Addison, J. T. and Portugal, P. (2003). Unemployment duration: Competing and defective risks. *Journal of Human Resources*, 38: 156–191.

Andersen, P. K. and Gill, J. P. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10: 1100–1120.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, Netherlands: Kluwer Academic.

Beck, N., Katz, J. and Tucker, R. (1998). Taking time seriously: Time-series cross-section analysis with a binary dependent variable. *American Journal of Political Science*, 42: 1260–1288.

Berry, F.S. and Berry, W.D. (1990). State lottery adoptions as policy innovations: An event history analysis. *American Political Science Review*, 84: 573–579.

Blossfeld, H. and Rohwer, G. (2001). *Techniques of Event History Modeling*, 2nd edn. Nahwah, NJ: Lawrence Erlbaum.

Box-Steffensmeier, J. and DeBoef, S. (2006). Repeated events survival models: The conditional frailty model. *Statistics in Medicine*, 25: 3518–3533.

Box-Steffensmeier, J. and Jones, B. S. (2004). *Event History Modeling: A Guide for Social Scientists*. Cambridge, UK: Cambridge University Press.

Box-Steffensmeier, J. and Zorn, C. (2001). Duration models and proportional hazards in political science. *Amerian Journal of Political Science*, 45: 972–988.

Collett, D. (1994). *Modeling Survival Data in Medical Research*, 1st edn. New York: Chapman & Hall.

Cook, R. J., Fleishman, E. and Grimes, V. (1991). *A National Evaluation of Title IV-E Foster Care Dependents Living Programs for Youth in Foster Care: Phase 2, Final Report, Volume 7*. Rockville MD: Westat.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society*, B, 30: 248–275.

Darmofal, D. and Young, L. (2006). Bayesian spatial survival models in political science. Presented at the Annual Meeting of the Political Science Association, Philadelphia, PA.

Diermeier, D. and Stevenson, R. (1999). Cabinet survival and competing risks. *American Journal of Political Science*, 43: 1051–1068.

Fanshel, D., Finch, S. J. and Grundy, J. F. (1990). *Foster Children in a Life Course Perspective*. New York: Columbia University Press.

Gifford, E. and Foster, E. M. (2005). Provider-level influences on receipt of aftercare services: A multilevel hazard model (No. 00-77). University Park, PA: Pennsylvania State University, Methodology Center.

Goldstein, J., Freud, A. and Solnit, A. J. (1973). *Beyond the Best Interests of the Child*. New York: Free Press.

Golub, J. (forthcoming). Event history analysis. In J. M. Box-Steffensmeier, H. Brady and D. Collier (eds), *Oxford Handbook on Political Methodology*. New York: Oxford University Press.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81: 515–526.

Harrell, F. (1986). The PHGLM procedure SAS supplementary library user's guide, version 5. Cary, NC: SAS Institute.

Hosmer, D. W. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Keele, L. (2005). Covariate functional form in Cox models. Unpublished manuscript.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis : Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

Leiberman, A. F. (1987). Separation in infancy and early childhood: Contributions of attachment theory and psychoanalysis. In J. B. F. a. S. Bloom-Feshbach (ed.), *The Psychology of Separation and Loss: Perspectives on Development, Life Transitions, and Clinical Practice*. San Francisco: Jossey-Bass.

Lin, D. Y., Wei, L. J. and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika*, 85: 605–618.

Oneal, J. and Russett, B. (1997). The classical liberals were right: Democracy, interdependence, and conflict. *International Studies Quarterly*, 41: 267–294.

Royston, Patrick. (2006). Explained variation for survival models. *STATA Journal*, 6: 83–96.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis*. New York: Oxford University Press.

Teachman, J. and Hayward, M. (1993). Interpreting hazard rate models. *Sociological Methodology and Research,* 21: 340–371.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data*. New York: Springer-Verlag.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77: 147–160.

Volden, Craig (2006). States as policy laboratories: Emulating success in the Children's Health Insurance Program. *American Journal of Political Science*, 50: 294–312.

Zimmerman, R. B. (1982). *Foster Care in Retrospect (Studies in Social Welfare)*. New Orleans: School of Social Work, Tulane University.

Zorn, C. (2000). Modeling duration dependence. *Political Analysis*, 8: 367–380.

This page intentionally left blank

Chapter 26

Parametric event history analysis: an application to the analysis of recidivism

Hee-Jong Joo

1 Introduction

Event history (or survival analysis[1]) has been developed and used for the analysis of longitudinal data on the occurrence of events (see Allison, 1984; Namboodiri and Suchindran, 1987; Schmidt and Witte, 1988; Blossfeld, Hamerle and Mayer, 1989; Yamaguchi, 1991; Singer and Willett, 2003; Box-Steffensmeier and Jones, 2004). It is a general methodology for studying a transition from one event (or state) to another, and research interest centers on whether and, if so, when events occur. In event history analysis, we can parameterize both the probability of an event's occurrence and the timing of the event for those who will ultimately experience the event. This method is appropriate when the variable of interest is the time interval (e.g., years, months, days, or seconds) between the initial event (e.g., marriage) and a subsequent terminal event (e.g., divorce).

Social scientists are interested in various kinds of events and concerned with the patterns and correlates of the occurrences of events (Yamaguchi, 1991). An event is made up of some "qualitative change" that happens at a specific point in time. For work and career researchers, for instance, job changes, promotions, layoffs, and retirements can be thought of as major events. Demographers study births, deaths, marriages, divorces, and migrations. The major events of interest in criminological studies include crimes, arrests, convictions, and incarcerations. In event history analysis, event occurrence is defined in terms of an individual's transition from one state to another. In a recidivism study, for example, the first state is "released from prison" and the second state is "returned to prison." The term "survival" here describes a continuation of the state of being "released from prison" and is thus the opposite of recidivism. The survival probabilities then can be defined as the cumulative proportion surviving at the end of a specified time interval (e.g., month or year) or 1 minus the recidivism rate.

Another key concept in event history analysis is the hazard rate or hazard function. The hazard rate, h(t), is the probability that persons who

[1] Event history analysis is often referred to as survival analysis. In biomedical and engineering research, for example, much of the literature on event history methods goes by the name of survival analysis. It is also referred to as duration models, failure-time models, and reliability models.

did not experience an event (e.g., reincarceration) at the beginning of a specified time interval (e.g., month) will experience the event (e.g., return to prison) during that interval, given that the individual is at risk at that time. In each time interval, the hazard rate can be calculated by dividing the number of events by the number of individuals at risk. The hazard rate, which is an unobserved variable and usually varies with time and among groups, is the "fundamental" dependent variable in an event history model. For most applications, however, hazard rate is re-expressed by 1/h(t) so that the dependent variable is transformed into the expected length of time until an event occurs (T or Log T). One of the most important characteristics of the hazard rate is that "it controls both the occurrence and the timing of events" (Allison, 1984, p. 16).

This statistical method allows researchers to examine not only individuals' survival or failure in terms of survival probability or hazard rate, but also the relationship between the length of the survival time and independent variables, or covariates, of theoretical interest (Box-Steffensmeier and Jones, 2004). Event history analysis, for example, enables criminologists to examine not only whether or not an individual was rearrested during a certain follow-up period (whether events occur), but also in how many days after release from prisons released parolees were rearrested (when events occur), with some attention to the factors related to the probability and the timing of an event. Event history analysis thus provides a different way of viewing recidivism and allows comparison among various groups. As Allison (1984) argues, the best way to study events and their cause is to collect event history data and examine the patterns and correlates of the occurrences of events with event history analysis. To determine whether a research question calls for event history analysis, it is helpful to conduct the above-mentioned "whether and when test" (Singer and Willett, 2003, p. 306).

2 Problems of conventional methods in the analysis of event history data: recidivism as an example

In the past, recidivism has been, in most cases, reported merely as the percentage of parole releasees who returned to prison within a certain period of follow-up. However, the prediction of parole outcome is not limited simply to success or failure. In many cases, there exist substantial differences in the timing of recidivism across demographic and criminal behavior characteristics of each recidivist. Thus, the timing of an event of interest is recognized as an important factor in the recidivism study.

In previous studies of recidivism (e.g., Rossi et al., 1980), however, ordinary multiple regression methods were applied to event history data. The events of interest were either arrests, convictions, or incarcerations, and the aim was to determine how the probability of an event depends on several explanatory variables such as age at release, gender, race/ethnicity, education, and prior criminal behavior. The dependent variable is a dummy variable indicating whether or not an individual was rearrested (or reconvicted or reincarcerated) during a certain follow-up period (12 months in the Rossi et al. study). As Allison (1984, pp. 10–11) pointed out, however, this method is still not ideal:

> Aside from the well-known problems in the use of ordinary least squares with a dummy dependent variable (Hanushek and Jackson, 1977, Ch. 7), dichotomizing the dependent variable is arbitrary and wastes information. It is arbitrary because there was nothing special about the 12-month dividing line except that the study ended at that point. Using the same data, one might just as well compare those arrested before or after the six-month mark. It wastes information because it ignores the variation on either side of the dividing line. One might suspect, for example, that someone arrested immediately after release had a higher propensity toward criminal activity than someone arrested 11 months later.

In an effort to avoid these problems, the length of time from release to arrest or return to prison will be used as the dependent variable in a multiple regression. However, a substantial proportion of each cohort did not return to prison during a certain follow-up period, which is censored. Censoring exists when incomplete information is available about the duration of the risk period due to a limited time of observation. Exclusion of the censored cases can lead to severe bias or loss of information in the parameter estimates of conventional statistical procedures such as ordinary least square (OLS) regression. As an alternative solution, one might assign the maximum length of time observed as the value of the dependent variable for the censored cases. But this obviously underestimates the true value and substantial bias may result (Allison, 1984; Yamaguchi, 1991).

In addition to this censoring problem, there are also difficulties incorporating time-varying explanatory variables in a multiple regression which predicts timing of an event (Sorensen, 1977; Tuma and Hannan, 1978; Allison, 1984; Ekland-Olson et al., 1991). If a study intends to examine possible causes of events or to determine how the probability of an event depends on several explanatory variables, the event history should also include data on possible time-varying explanatory variables. While some explanatory variables, such as race and sex, are constant over time, other variables (e.g., income, marital status, and age) may vary with time. To avoid these two typical problems—censoring and the mishandling of time-varying explanatory variables—with the conventional multiple regression approach to event history data, event history analysis has been used in many areas of the social and behavioral sciences.

3 Parametric versus nonparametric event history methods

Within event history analysis, there are several approaches to analyzing the data. They include distribution versus regression and parametric versus nonparametric methods.

3.1 Distribution versus regression methods

The distribution method—sometimes linked to life table analysis—examines the distribution of time until an event occurs, or the time between events. This is one of the most common methods applied in demographic studies, and it is basically nonparametric in nature.[2] In this method, survival times are typically measured at monthly (or other) intervals permitting the computation of detailed survival trajectories for each cohort.

Two related functions are used in the analysis: (1) the survival probabilities—the cumulative proportion surviving at the end of a specified time interval, or 1 minus recidivism rate; and (2) the hazard rate—the probability that persons who did not experience an event at the beginning of a specified time interval will experience the event during that interval. It is the rate of the occurrence of the event during the risk period (see Figure 26.1).

Much of the early work on event history analysis in the biomedical area, and the intra- and inter-cohort analysis of recidivism rates in criminology, can be included in this category (Ekalnd-Olson et al., 1991; Joo, 1993). In Joo's (1993) inter-cohort recidivism study, for example, this research design allowed for a 36-month follow-up for all four release cohorts to determine not only if parolees were reincarcerated but, if so, how long after release they returned to prison. The life table method examines the pace at which the offenders recidivate at monthly

[2] "Nonparametric statistics are designed to be used when the data being analyzed depart from the distributions that can be analyzed with parametric statistics. In practice, this most often means data measured on a nominal or an ordinal scale. Nonparametric tests generally have less power than parametric tests. The chi-square test is a well-known example" (Vogt, 1999, p. 192).

Intvl start time	*Number entering this intvl*	*Number wdrawn during intvl*	*Number exposd to risk*	*No. of termnl events*	*Propn terminating*	*Propn surviving*	*Cumul propn surv at end*	*Probability density*	*Hazard rate*	*SE of cumul surviving*	*SE of probability density*	*SE of hazard rate*
0.0	1199.0	0.0	1199.0	0.0	0.0000	1.0000	1.0000	0.0000	0.0000	0.000	0.000	0.000
1.0	1199.0	0.0	1199.0	0.0	0.0000	1.0000	1.0000	0.0000	0.0000	0.000	0.000	0.000
2.0	1199.0	0.0	1199.0	0.0	0.0000	1.0000	1.0000	0.0000	0.0000	0.000	0.000	0.000
3.0	1199.0	0.0	1199.0	0.0	0.0000	1.0000	1.0000	0.0000	0.0000	0.000	0.000	0.000
4.0	1199.0	0.0	1199.0	0.0	0.0000	1.0000	1.0000	0.0000	0.0000	0.000	0.000	0.000
5.0	1199.0	0.0	1199.0	0.0	0.0000	1.0000	1.0000	0.0000	0.0000	0.000	0.000	0.000
6.0	1199.0	0.0	1199.0	0.0	0.0000	1.0000	1.0000	0.0000	0.0000	0.000	0.000	0.000
7.0	1199.0	0.0	1199.0	33.0	0.0275	0.9725	0.9725	0.0275	0.0279	0.005	0.005	0.005
8.0	1166.0	0.0	1166.0	23.0	0.0197	0.9803	0.9533	0.0192	0.0199	0.006	0.004	0.004
9.0	1143.0	0.0	1143.0	19.0	0.0166	0.9834	0.9374	0.0158	0.0168	0.007	0.004	0.004
10.0	1124.0	0.0	1124.0	41.0	0.0365	0.9635	0.9033	0.0342	0.0372	0.009	0.005	0.006
11.0	1083.0	0.0	1083.0	26.0	0.0240	0.9760	0.8816	0.0217	0.0243	0.009	0.004	0.005
12.0	1057.0	0.0	1057.0	21.0	0.0199	0.9801	0.8641	0.0175	0.0201	0.010	0.004	0.004
13.0	1036.0	0.0	1036.0	14.0	0.0135	0.9865	0.8524	0.0117	0.0136	0.010	0.003	0.004
14.0	1022.0	0.0	1022.0	24.0	0.0235	0.9765	0.8324	0.0200	0.0238	0.011	0.004	0.005
15.0	998.0	0.0	998.0	26.0	0.0261	0.9739	0.8107	0.0217	0.0264	0.011	0.004	0.005
16.0	972.0	0.0	972.0	22.0	0.0226	0.9774	0.7923	0.0183	0.0229	0.012	0.004	0.005
17.0	950.0	0.0	950.0	17.0	0.0179	0.9821	0.7781	0.0142	0.0181	0.012	0.003	0.004
18.0	933.0	0.0	933.0	30.0	0.0322	0.9678	0.7531	0.0250	0.0327	0.012	0.005	0.006
19.0	903.0	0.0	903.0	27.0	0.0299	0.9701	0.7306	0.0225	0.0304	0.013	0.004	0.006
20.0	876.0	0.0	876.0	18.0	0.0205	0.9795	0.7156	0.0150	0.0208	0.013	0.004	0.005
21.0	858.0	0.0	858.0	7.0	0.0082	0.9918	0.7098	0.0058	0.0082	0.013	0.002	0.003
22.0	851.0	0.0	851.0	19.0	0.0223	0.9777	0.6939	0.0158	0.0226	0.013	0.004	0.005
23.0	832.0	0.0	832.0	13.0	0.0156	0.9844	0.6831	0.0108	0.0157	0.013	0.003	0.004
24.0	819.0	0.0	819.0	11.0	0.0134	0.9866	0.6739	0.0092	0.0135	0.014	0.003	0.004

Figure 26.1 Life table: 24 months follow-up period

intervals. In this sense, the design is longitudinal, providing for the evaluation over time of inmates released in a given year as well as the assessment of changes across cohorts of parolees. The life table method, however, does not permit the use of covariates or explanatory variables in the model. In the example to be presented below, however, the focus is on regression-like methods in which the occurrence of events (or hazard rate) is dependent on a linear function of explanatory variables. This method with regard to recidivism has become popular in the study of recidivism (e.g., Farrington and Tarling, 1985; Schmidt and Witte, 1987 and 1988; Ekland-Olson et al., 1991; Joo, 1993). Schmidt and Witte (1988) point out the usefulness of survival models which include explanatory variables:

> The use of explanatory variables allows one to make statements about the way in which survival time is influenced by individual characteristics, criminal history, or structural variables, and it also allows one to make predictions for individuals and not just for random samples of releases.

As mentioned above, the hazard rate or h(t) is used as a dependent variable in this method, and it controls both the occurrences and the timing of events.

3.2 Parametric[3] versus nonparametric methods

Researchers have relied on a variety of model specifications to predict time until recidivism. There are two major groups of methods for analyzing hazard rates: parametric and nonparametric methods. Parametric methods estimate the effects of explanatory variables, or covariates, on hazard rates. We can analyze both "time-invariant" covariates, which do not vary throughout the duration of observation (e.g., race, gender, and age at first arrest) and "time-varying" explanatory variables (e.g., age, income, and marital status) in parametric models.

On the other hand, nonparametric methods do not specify the relation between hazard rates and covariates. Instead, separate estimates of hazard rates as a function of time (e.g., hazard rates calculated in a monthly interval) are obtained for each group of time-invariant categorical variables such as gender and race (Yamaguchi, 1991, p. 3). In most criminological studies, as Schmidt and Witte (1988, p. 18) pointed out, parametric models are more commonly used and, "fit recidivism data better and provide more accurate predictions than nonparametric models do."

While no specific distribution is assumed for the time until recidivism in nonparametric methods, parametric methods assume a certain type of distribution in which the hazard rates depend on time. One of the advantages of parametric event history models lies in their capability to directly model the time dependency exhibited in event history data, and this can be done by specifying a distribution function for the hazard rate (Box-Steffensmeier and Jones, 2004). When there are strong theoretical expectations or previous empirical findings regarding the shape of the hazard rate, parametric models would be most reasonable. If the researcher knows or suspects that the hazard rate increases or decreases over time, then one may specify a distribution that reflects such a relationship. By correctly specifying the shape of the hazard rate, the researcher can obtain better estimates of the time dependency in the data as well as more precise estimates of covariate parameters (Box-Steffensmeier and Jones, 2004, p. 21).

[3] Parametric statistics are "statistical techniques designed for use when data have certain characteristics—usually when they approximate a normal distribution and are measurable with interval or ratio scales. Also statistics used to test hypotheses about population parameters" (Vogt, 1999, p. 206).

Of a variety of parametric models, the selection of a particular parametric model depends primarily on the shape of the hazard rate function [h(t) or log h(t)]. In the process of model selection, as Allison (1984, pp. 30–31) suggests, "first choice is between the exponential regression model, in which there is no dependence on time, and all others.... If the exponential model is rejected, one must then choose between monotonic models (in which the hazard always increases or always decreases with time) and nonmonotonic models (in which the hazard may sometimes increase and sometimes decrease)."

The exponential model assumes that the baseline hazard rate is flat with respect to time. This means that the hazard rate is constant or the same at all points in time. We can express hazard rate for the exponential distribution as

$$h(t) = \lambda, \quad (t > 0, \lambda > 0) \qquad (1)$$

where λ is a positive constant. On the other hand, the baseline hazard of monotonic models such as the Weibull model can monotonically increase, monotonically decrease, or remain constant with respect to time. To show how the Weibull distribution can monotonically vary with time, the hazard rate for the Weibull model is given by

$$h(t) = \lambda p(\lambda t)^{p-1}, \quad (t > 0, \lambda > 0, p > 0) \qquad (2)$$

where λ is a positive constant and p is the shape parameter, which determines the shape of the hazard rate. When $p > 1$, for example, the hazard rate monotonically increases with time; when $p < 1$, the hazard rate monotonically decreases with time; when $p = 1$, the hazard is a constant value λ, which is flat (Box-Steffensmeier and Jones, 2004, p. 25). The exponential model thus can be considered a special case of the Weibull model.

While the Weibull model permits the hazard to change in one direction with respect to time, the assumption of the monotonic hazard rate sounds somewhat unrealistic for many research questions in social science research. In many cases, there are good reasons to suspect that the hazard rate changes autonomously with time. In fact, most survival models in criminology have a nonmonotonic hazard rate, in particular a hazard rate that first rises and then falls.

The log-normal and log-logistic models assume a nonmonotonic hazard distribution in which hazard first increases, reaches a peak, and then gradually declines (Allison, 1984; Schmidt and Witte, 1988; Yamaguchi, 1991; Joo et al., 1995; Box-Steffensmeier and Jones, 2004). Since this is a pattern found in most of the recidivism data in criminology (see Figure 26.2), we rely on these two specifications for the multivariate prediction of survival time, which will follow in the next section.

Log-normal and log-logistic models are specific cases of a general class of models known as accelerated failure time models. If T is the mean

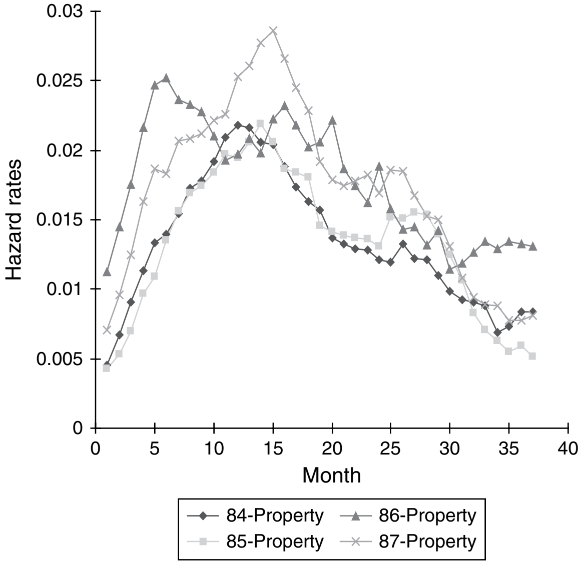


Figure 26.2 Hazard rates for the 1984–1987 property cohorts

Source: Joo et al. (1995). Recidivism among paroled property offenders released during a period of prison reform. *Criminology*, Vol.33 (No.3).

(survival) time until recidivism, these models may be written as

$$\text{Log } T = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots b_kx_k + e \quad (3)$$

where e is a random disturbance term that depends on time and explanatory variables. According to the differences in disturbance term e, each specific model is determined. Commonly assumed distributions include extreme values, normal, and logistic distributions, which give rise to corresponding Weibull, log-normal, and log-logistic distributions for the mean time until recidivism (Allison, 1984, p. 30).

As Blossfeld et al. (1989, p. 249) point out, "the log-logistic model, along with the log-normal distribution, are the most commonly recommended distributions, if an initial increasing and then a decreasing risk is presumed to exist." In a recidivism study, which will follow in the next section as an example of empirical study utilizing parametric event history analysis, a comparison will be made between the two models to determine which one fits the data better (see Schmidt and Witte (1988) and Box-Steffensmeier and Jones (2004) for a more detailed discussion of log-normal and log-logistic models).

The aim of the following study is to estimate an appropriate prediction model of the mean time until reincarceration to determine how the survival time depends on individual parolees' characteristics, such as age at release, gender, race/ethnicity, risk assessment score, and prior incarceration offense. These models also allow us to estimate the effect of individual characteristics on the time until recidivism. For example, as Chung et al. (1991, p. 60) illustrate, "if any of the parolee characteristics is a dummy variable indicating participation or nonparticipation in some correctional program," using survival time model, we can estimate the effect of a certain correctional program on the length of time until recidivism by properly controlling for the other explanatory variables in the equation. Because the data show a nonmonotonic pattern for the hazard rates, log-normal, and log-logistic models are computed and compared in order to estimate appropriate multivariate prediction models of the mean time until recidivism. The aim is to illustrate the estimation of how the survival time depends on parolees' characteristics, with some attention to the identification of factors related to the probability and the timing of parole failure.

4 Multivariate prediction of survival time: an example of log-normal and log-logistic event history analysis

4.1 Introduction

With regard to prison overcrowding and ensuing judicial and political pressures on ever-worsening prison conditions, there has been increasing interest in finding ways to use the limited available prison space in the most effective manner for controlling crime. Many studies have indicated that a relatively small group of offenders are responsible for a disproportionately high volume of criminal activity. In this respect, selective incapacitation, as a strategy for achieving crime control, has gained increased attention. Central to this policy is the problem of predicting an individual's criminality, as well as various ethical and policy issues posed by this prediction and resulting policy (Greenwood, 1982; Greenwood and Abrahamse, 1982; Cohen, 1983; Blumstein, Cohen, Martin and Tonry, 1983; von Hirschi and Gottfredson, 1983–84; Klein and Caggiano, 1986; Blumstein, Cohen, Roth and Visher, 1986; Tonry, 1987; Gottfredson and Tonry, 1987; Haapanen, 1990; Joo et al., 1995; Petersilia, 2003).

Selective incapacitation involves "individually based" sentences. Such sentences would vary with differences in predictions of the individual propensity to commit future crimes.

Therefore, the effectiveness of a selective incapacitation policy in reducing crime depends on the capability to identify those offenders who continue to threaten public safety. As Ekland-Olson et al. (1991, p. 101) also point out, "as states continue to rely on parole as a means of easing prison crowding and as political pressure accumulates regarding associated public safety risks, prediction of who is more or less likely to recidivate becomes increasingly important."

In event history analysis, as mentioned above, we can parameterize both the probability of eventual failure and the distribution of survival times for those who will ultimately fail. In a previous study (Joo, 1993), we reconfirmed commonly found correlates of the recidivism by survival analysis (distribution or life-table method in particular). That is, the knowledge about an individual's risk assessment score, race/ethnicity, age, and release offense were found to be very important factors in predicting the probability of returning to prison among property offenders across cohorts. Some variations were also found in the probability of reincarceration across individual parolees' characteristics.

However, the prediction of parole outcome is not limited to success or failure. The trends in the hazard rates in Joo et al. (1995) indicate that there are important differences in the timing of reincarceration across parolee characteristics, as well as across cohorts. The objective here is to construct and estimate an appropriate multivariate prediction model of the length of the time from release to return to prison (or survival time), with attention to how the factors that explain parole outcome also are related to the timing of reincarceration among paroled property offenders.

In our previous study (Joo, 1993), which examined both intra-cohort variation in reincarceration and inter-cohort comparison of reincarceration rates, our focus was on the overall probabilities and the shifting patterns of reincarceration during this period of dramatically revised criminal justice policies. In addition, we compared a cohort of inmates released under the Prison Management Act (PMA) with a comparable release cohort (i.e., 1987 property cohort) consisting of inmates who did not receive accelerated release under the provision of the PMA. The aim of this comparison was to examine if PMA releasees differ from non-PMA releasees in recidivism pattern. The primary analytic strategy was to compare the three-year survival patterns of four successive parole cohorts through the use of the distribution (or life-table) method.

In multivariate prediction of survival time, however, our focus is placed on regression-like models in which the occurrence of an event depends on several explanatory variables, such as parolees' characteristics and prior criminal history. We use these models to make predictions of the mean time until reincarceration, and to estimate the effect of individual characteristics on survival time.

4.2 Analysis of survival time

The survival time model (regression-like method here) focuses on those who returned to prison and examines the length of time from release to reincarceration. The aim is to estimate how the survival time depends on individual parolees' characteristics. They include age at release, gender, race/ethnicity, assessed risk, and prior incarceration offense. All of these explanatory variables are constant in value over the follow-up period. The dependent variable in this analysis is the length of time from release on parole to return to prison and includes censored observations.

To simplify the interpretation of the model, we recoded some of the explanatory variables in the analysis. Prior incarceration offense (PIO) for paroled property offenders was collapsed into four property categories—burglary, larceny/theft, fraud/forgery, and motor vehicle theft. Age was divided into three categories—18 to 27, 28 to 37, and 38 and older. The coding of gender, race/ethnicity (Anglo, African-American, and Latino) and risk assessment score (high, medium, and low) were not altered. Parolees who did not fail during the 36-month follow-up period are considered censored. Similar to dummy variable regression, each variable has a reference group. In the analyses that follow, the reference groups are: 38+, low risk, "fraud/forgery" prior offense, female, and Anglo.

As mentioned above, the selection of a particular parametric model depends primarily on the distribution of the hazard rates. In this connection, the log-normal and log-logistic models assume a nonmonotonic hazard distribution, specifically one that initially increases and then declines. Since this is the same pattern found in the hazard distribution of our data, we fit the two distributions to the data, by maximum likelihood method, to determine which one fits our data better.

Variation in survival time among four successive cohorts

Table 26.1 reports estimates for both log-normal and log-logistic models for the two early release cohorts: coefficient (b), anti-log of b, and the t ratio. The coefficient estimates (b) are like unstandardized regression coefficients in that they depend on the metric of each explanatory variable. A negative coefficient indicates that an increase in the corresponding element in X_i decreases the mean survival time until recidivism. Since the coefficients are in the log of length of time, the anti-logs have a more straightforward interpretation, i.e., the antilog of the coefficient represents the proportion of time from release to return compared to the reference group. For example, the antilog coefficient for high risk for the 1984 cohort (.65) indicates that high-risk parolees survived only 65% of the length of time that low-risk parolees survived. The asymptotic t ratio for each coefficient is calculated by dividing the parameter estimate by its asymptotic standard error. Like standardized regression coefficients (ß), these t statistics, under the null hypothesis that each coefficient is zero, are metric-free and give some indication of the relative importance of the explanatory variables. Also presented are the log-likelihood values and the significance of the variables in the models.

The overall results from these two models for the 1984 cohort are very similar. The likelihood values are −775.6 for the log-normal model and −784.4 for the log-logistic model. These likelihood values indicate that the log-normal model fits the data slightly better than the log-logistic model. The results from the estimation of both log-normal and log-logistic regressions for the 1984 cohort indicate that release offense, race/ethnicity, and assessed risk have significant net effects on the survival time.

The effect of prior incarceration offense is highly significant. Parolees for prior auto-theft conviction have a survival time that, all else equal, is 52% that of parolees for the reference group of forgery/fraud. Parolees with the prior incarceration offense of burglary have a survival time that is about 68% of parolees with the prior offense of forgery/fraud. The significant effects for motor-vehicle theft and burglary are not limited to the comparison of burglary and forgery/fraud, and auto theft and forgery/fraud. The difference between motor-vehicle theft and burglary is also significant.

The effects of race/ethnicity confirm our earlier findings that Anglo parolees survive longer than minorities. Both African-American and Latino parolees released in 1984 had survival times that were approximately 78% and 72% of the survival time, respectively, experienced by Anglos. These results indicate that Latino parolees had the lowest expected survival times among paroled property offenders.

As expected, high-risk parolees have a survival time that is 65% that of low-risk parolees. Those with medium-assessed risk have a survival time that is about 75% of that for the category of low risk. Although gender is not found as an important predictor

Table 26.1 Estimates for log-normal and log-logistic models predicting the possibility of recidivism: 1984 and 1985 cohorts

Characteristics	*1984 cohort*						*1985 cohort*					
	Log-normal model			*Log-logistic model*			*Log-normal model*			*Log-logistic model*		
	b	*Anti-log*	*t*	*b*	*Anti-log*	*t*	*b*	*Anti-log*	*t*	*b*	*Anti-log*	*t*
Constant	4.78	118.75	16.31	4.78	118.63	16.19	5.36	213.15	16.40	5.32	203.36	16.17
Race/Ethnicity	*			**			**			**		
African-American	−.25 *	.78	−2.36	−.28 **	.76	−2.63	−.51 **	.60	−4.34	−.52 **	.60	−4.54
Latino	−.33 *	.72	−2.37	−.35 *	.71	−2.52	−.37 *	.69	−2.53	−.39 **	.68	−2.73
Anglo												
Gender							**			*		
Male	−.24	.79	−1.20	−.27	.76	−1.33	−.66 **	.52	−2.62	−.60 *	.55	2.40
Female												
Age							*			**		
18–27	.07	1.08	0.48	.09	1.09	0.57	−.23	.80	−1.42	−.26	.63	−1.65
28–37	.18	1.20	1.14	.16	1.17	1.04	−.44 **	.65	−2.73	−.47 **	.81	−2.97
38+												
Risk score	*			*			*			*		
High	−.43 *	.65	−2.31	−.42 *	.66	−2.26	−.47 *	.63	−2.55	−.47 **	.63	−2.59
Medium	−.29 *	.75	−2.01	−.31 *	.74	−2.07	−.19	.83	−1.31	−.21	.81	−1.45
Low												
Release offense	**			**								
Burglary	−.39 *	.68	−2.45	−.37 *	.69	−2.32	−.18	.83	−1.09	−.17	.84	−1.06
Larceny/Theft	−.17	.85	−.97	−.14	.87	−.78	.13	1.14	.72	.15	1.17	.85
MVT	−.66 **	.52	−3.01	−.64 **	.53	−3.00	−.33	.72	−1.41	−.34	.71	−1.54
Forgery/Fraud												
Scale	1.17			.67			1.10			.62		
Log-likelihood		−775.6			−784.4			−589.8			−594.3	

*Significant at .05 level.
**Significant at .01 level.

of survival time in this cohort, males survive approximately 79% of the length of time female parolees survive.

As mentioned above, the asymptotic t ratios are useful statistics for moderate to large samples. If the ratio exceeds 2, the coefficient is significantly different from zero at the .05 level with a two-tailed test. Besides, the relative sizes of these ratios can be used to measure relative importance of the variables, given that all predictors are in the model. In this cohort, we see that release offense, race/ethnicity, and risk score have significant effects on the time until recidivism. Particularly, prior incarceration offense of motor-vehicle theft is the most significant predictor (−3.0) in predicting the length of time from release to reincarceration, followed by burglary (−2.45), Latino (−2.37), African-American (−2.35), and high risk (2.31). In this regard, the type of parolee most likely to have a small value of survival time until reincarceration is the Latino auto thief who is assessed as high risk.

The models for the 1985 release cohort show a somewhat different pattern of effects from the 1984 models. Overall, the log-normal model fits the 1985 data slightly better than log-logistic model. The likelihood values are −589.76 and −594.26 for the log-normal model and the log-logistic model, respectively. Race/ethnicity has the most significant effect on survival time ($t = -4.34$ for the log-normal model). African-American and Latino parolees survived only 60% and 69% of the length of time, respectively, that Anglo parolees survived. The differences between African-Americans and Latinos are also statistically different. Gender differences in return time are quite strong. Males have a survival time that averages about 52% of that for female parolees. The gender effect in this model is considerably stronger than that reported for the previous cohort. Assessed risk also has a significant effect on return time. Parolees with high-assessed risk experienced a survival time that is only 63% of that for those with low risk. The comparable proportion for medium-risk releasees is 83%. The difference between high and medium risk is also significant.

Finally, the effects of age are also significant—parolees aged 28 to 37 return more quickly compared to those 18 to 27 and 38 and over. Those aged 28 to 37 had a survival time that is only 65% of that for 38 and over. Parolees 18 to 27 have also lower expected survival time compared to the oldest age group. However, there are no significant differences between the 18–27 and 38+ age groups. For the 1985 cohort, the type of parolee most likely to return to prison shortly after release was the 28–37-year-old African-American male who was assessed as high risk.

Table 26.2 also presents estimates for both log-normal and log-logistic models for the two later release cohorts. For the 1986 release cohort, the results from the log-normal model are also very similar to the corresponding results from the log-logistic model (the likelihood values are −1044.5 for the log-normal model and −1055.3 for the log-logistic model). However, the patterns of effects for both models are somewhat different from the previous models in the two previous cohorts. The three significant predictors for the 1986 cohort are risk assessment score, race/ethnicity, and age.

Once again, the effects of risk are the strongest ($t = -5.4$ for high risk for both models). Those with high-assessed risk have a net survival time that is 43% of that for parolees with low-assessed risk. The group with medium-assessed risk also differs significantly from the reference category of low risk. Moreover, those with high risk are significantly different from those with medium risk.

The effects of race/ethnicity are also very strong. The t ratios are −3.34 and −3.0 for Latino and African-American, respectively. As was the case for the 1984 cohort, Latino parolees showed shorter survival time until reincarceration compared to African-American

Table 26.2 Estimates for log-normal and log-logistic models predicting the possibility of recidivism: 1986 and 1987 cohorts

Characteristics	*1986 cohort*						*1987 cohort*					
	Log-normal			*Log-logistic*			*Log-normal*			*Log-logistic*		
	b	*Anti-log*	*t*	*b*	*Anti-log*	*t*	*b*	*Anti-log*	*t*	*b*	*Anti-log*	*t*
Constant	4.86	129.15	18.31	4.83	125.09	18.54	4.86	129.02	23.26	4.82	123.72	23.43
Race/Ethnicity	**			**			**			**		
African-American	−.29 **	.75	−3.00	−.32 **	.73	−3.34	−.24 **	.79	−2.92	−.24 **	.79	−2.97
Latino	−.41 **	.66	−3.34	−.45 **	.64	−3.67	−.01	.99	−.08	.00	1.00	−.04
Anglo												
Gender							*			*		
Male	−.24	.79	−1.27	−.20	.82	−1.12	−.33 *	.72	−2.33	−.33 *	.72	−2.40
Female												
Age	*			*			**			**		
18–27	−.33 *	.72	−2.32	−.34 *	.71	−2.40	−.64 **	.53	−4.92	−.65 **	.52	−5.05
28–37	−.12	.89	.80	−.12	.88	−.87	−.39 **	.68	−2.98	−.38 **	.68	−2.94
38+												
Risk score	**			**			**			**		
High	−.86 **	.43	−5.41	−.86 **	.43	−5.40	−.73 **	.48	−5.42	−.76 **	.47	−5.72
Medium	−.37 **	.69	−2.63	−.36 *	.70	−2.56	−.42 **	.66	−3.79	−.42 **	.66	−3.80
Low												
Release offense												
Burglary	−.23	.79	−1.54	−.23	.79	−1.58	.14	1.15	1.26	.18	1.19	1.57
Larceny/Theft	−.07	.93	−.43	−.06	.94	−.41	.06	1.06	.46	.07	1.07	.56
MVT	−.12	.89	−.59	−.10	.90	−.50	−.12	.88	−.79	−.08	.92	−.54
Forgery/Fraud												
Scale	1.18			0.69			1.10			0.63		
Log-likelihood	−1044.5			−1055.3			−1171.8			−1178.7		

*Significant at .05 level.
**Significant at .01 level.

parolees. They have an expected length of time from release to reincarceration that is just under 66% of that for the reference group of Anglos. African-American parolees (75%) also differ significantly from Anglos. However, the difference between African-Americans and Latinos is not statistically significant.

The parameter estimates for age indicate that parolees 18 to 27 have a significantly lower expected survival time compared to the oldest age group. That pattern also holds for the 28 to 37-year-old age group. Finally, the 18 to 27 and 28 to 37 age groups are significantly different as well. Males have an average survival time that is 79% of that for females. However, gender is not found to be an important predictor of survival time in this cohort. Differences between males and females in the time until recidivism is also considerably smaller than that reported for the 1985 cohort. For the 1986 cohort, young Latino parolees with assessed high risk have the highest likelihood of quickly returning to prison after release.

The pattern of effects for the 1987 cohort are almost identical to that for the 1985 model in that both cohorts have the same variables which estimated coefficient found to be significantly different from zero. Overall, the 1987 model is significant with a log-likelihood of −1172 for log-normal model and 1179 for log-logistic model.

High-risk parolees have a net survival time that is only 48% of that for those with low-assessed risk. Those with medium risk have a time from release to reincarceration that is 66% of the low-risk group. Finally, high- and medium-risk parolees are also statistically different.

Age differences in return time are quite strong. Parolees aged 18 to 27 have substantially lower expected survival time than other age groups. Parolees 18 to 27 have also significantly lower survival time than the oldest age group. The difference between 18 to 27 and 28 to 37 is also significant.

With regard to race/ethnicity, African-American parolees have significantly lower survival times compared to Latinos and Anglos. However, Latinos and Anglos are essentially the same with regard to length of time from release to return. The gender effect is important for this cohort. Males have a significantly lower average survival time compared to females (69%). As has been the case with the two previous (1984 and 1985) cohorts, prior incarceration offense in the 1987 cohort has a trivial effect on return time, net of the other variables in the model. The type of parolee in the 1987 cohort most likely to return quickly to prison is a young African-American male with the assessment score of high risk.

In sum, the effects of risk are the strongest and the most consistent across the cohorts. We also observed consistently significant effects of race/ethnicity and age. With regard to race/ethnicity, the one persistent effect is that between Anglos and minorities. In two cases, however, the differences between Latinos and African-Americans are also statistically significant. When significant, the effects of age were rather strong, indicating that not only do younger parolees have a greater likelihood of failure but also a shorter survival time after release. Unlike the violent offenders whose criminal inclination is declining after their late-20s or 30s, paroled property offenders aged 28–37 showed still relatively short mean time until recidivism. Prior incarceration offense is found to have a trivial net effect. One likely explanation for this result is that other characteristics, such as age and risk, are associated with prior offense and once these factors are partialled out, offense has no effect.

Following our previous study (Ekland-Olson et al., 1991; Joo, 1993), we compute predicted values for survival time since this is a useful way to summarize the differences across the four models. Based on the estimates of t ratios for the log-normal model, we selected the characteristics in each model that are statistically significant but also that are associated with the lowest survival times. Taking a worst case scenario, the prediction equations for the four cohorts consist of the following characteristics:

1984—auto thief, high risk, Latino
1985—28–37, high risk, African-American, male
1986—18–27, high risk, Latino
1987—18–27, high risk, African-American, male

To obtain predicted values for survival time, we multiply the intercepts by the anti-logs of the coefficients associated with the selected characteristics. The following equations were used to generate the predicted survival time (S) in months:

1984 cohort: S = 119 * .52 * .65 * .72 = 29.0 months
1985 cohort: S = 213 * .65 * .63 * .60 * .52 = 27.2 months
1986 cohort: S = 129 * .72 * .43 * .66 = 26.4 months
1987 cohort: S = 129 * .53 * .48 * .79 * .72 = 18.7 months

As indicated in Table 26.1 and Table 26.2, risk assessment score has the strongest impact on survival time. For the 1986 and 1987 cohorts, this conclusion is also based on a comparison of the size of the coefficient for high risk (e.g., anti-log of .43 and .48 for the 1986 and 1987 cohorts, respectively).

A useful way to illustrate the impact of risk is to generate predicted survival times by substituting the coefficient for medium risk for the coefficient for high risk, leaving all of the other coefficients unchanged. The difference in predicted values illustrates the impact of high risk on expected survival times. The following equations were used to generate these revised predicted values.

1984 cohort: S = 119 * .52 * .75 * .72 = 33.4 months
1985 cohort: S = 213 * .65 * .83 * .60 * .52 = 35.9 months
1986 cohort: S = 129 * .72 * .69 * .66 = 42.3 months
1987 cohort: S = 129 * .53 * .66 * .79 * .72 = 25.7 months

Changing from high risk to the modal category of medium risk has a substantial impact on survival time. This change produces the following excess number of predicted survival months: 4.5, 8.6, 15.9, and 7.0 months for the 1984, 1985, 1986, and 1987 cohorts respectively.

As indicated, the impact of high risk on the predicted time until recidivism is most evident for the 1986 cohort. Although the predicted survival time for the 1986 cohort is larger than expected due to the differences in the number of significant variables across cohorts, this result still supports our early finding on the compositional difference in high-risk category in the 1986 cohort.

Variation in survival time between PMA and 1987 cohorts

Table 26.3 presents parameter estimates for the PMA cohort, as compared to the 1987 non-PMA cohort, for both log-normal and log-logistic models. While the overall results from these two models for the PMA cohort are very similar, unlike the cases for four successive cohorts, the log-likelihood values indicate that the log-logistic model (−1089.8) fits the PMA data slightly better than the log-normal model (−1109.9). For the 1987 non-PMA cohort, the log-likelihood value from the latter (−1171.8) is a little higher than the former (−1178.7), however both models are very similar. For both cohorts, therefore, we use parameter estimates from the log-logistic model to calculate the survival time.

Table 26.3 Estimates for log-normal and log-logistic models predicting the possibility of recidivism: PMA and 1987 cohorts

Characteristics	*PMA cohort*						*1987 cohort*					
	Log-normal			*Log-logistic*			*Log-normal*			*Log-logistic*		
	b	*Anti-log*	*t*	*b*	*Anti-log*	*t*	*b*	*Anti-log*	*t*	*b*	*Anti-log*	*t*
Constant	5.54	254.68	16.79	5.27	194.42	17.89	4.86	129.02	23.26	4.82	123.72	23.43
Race/Ethnicity	**			**			**			**		
African-American	−.46 **	.63	−3.60	−.40 **	.67	−3.57	−.24 **	.79	−2.92	−.24 **	.79	−2.97
Latino	.01	1.01	0.07	.00	1.00	.01	−.01	.99	−.08	.00	1.00	−.04
Anglo												
Gender	*			*			*			*		
Male	−.52 *	.59	−2.30	−.44 *	.64	−2.18	−.33 *	.72	−2.33	−.33 *	.72	−2.40
Female												
Age	**			**			**			**		
18–27	−.58 **	.56	−2.73	−.51 **	.60	−2.69	−.64 **	.53	−4.92	−.65 **	.52	−5.05
28–37	−.67 **	.51	−3.12	−.59 **	.55	−3.06	−.39 **	.68	−2.98	−.38 **	.68	−2.94
38+												
Risk score	**			**			**			**		
High	−1.20 **	.30	−5.79	−1.14 **	.32	−6.23	−.73 **	.48	−5.42	−.76 **	.47	−5.72
Medium	−.39 *	.68	−2.24	−.43 **	.65	−2.73	−.42 **	.66	−3.79	−.42 **	.66	−3.80
Low												
Release offense												
Burglary	−.28	.76	−1.61	−.22	.80	−1.41	.14	1.15	1.26	.18	1.19	1.57
Larceny/Theft	−.37 *	.69	−2.02	−.23	.79	−1.42	.06	1.06	.46	.07	1.07	.56
MVT	−.23	.79	−1.09	−.18	.84	−.94	−.12	.88	−.79	−.08	.92	−.54
Forgery/Fraud												
Scale	1.50			0.78			1.10			0.63		
Log-likelihood		−1109.9			−1089.8			−1171.8			−1178.7	

*Significant at .05 level.
**Significant at .01 level.

The results for the PMA cohort are identical to the results for the non-PMA cohort in terms of statistically significant variables. Both log-normal and log-logistic regressions for the PMA cohort indicate that risk assessment score, race/ethnicity, age, and gender have significant net effects on the survival time. However, noteworthy in this connection is that there are important differences in terms of the magnitude of the effects. Recall that the interpretation of the anti-log (exponentiated coefficients) is in terms of the proportion of survival time for a particular group compared to the reference group.

The most dramatic difference between the PMA and 1987 cohorts is with regard to assessed risk. High-risk PMA releasees have, all else being equal, a survival time that is only 32% of that of low-risk PMA parolees. The comparable figure for the 1987 group is 47%. The effect of race/ethnicity is somewhat stronger for PMA releasees as is the effect of gender. African-American and male PMA releasees have a lower expected survival time compared to their non-PMA counterparts. While PMA releasees aged 28–37 showed lower expected survival time than the 18–27 group, for the 1987 cohort, the youngest age group exhibited the lowest survival time.

Once again, as a useful way to summarize the differences between these two cohorts, we estimate the expected (or mean) survival time for parolees with characteristics that are statistically significant but also are associated with a high likelihood of return (i.e., for PMA cohort, high risk, male, and African-American, and being 28 to 37 years of age). The prediction equation for the 1987 cohort consists of the following characteristics: high risk, male, and African-American, and being 18 to 27 years of age. The predicted mean survival time (mean time from release to return) is obtained by adjusting the intercept by multiplying it times the coefficient for each of these characteristics. The following equations were used to generate the predicted survival time (S) in months:

PMA cohort: S = 194 * .32 (high) * .64 (male) * .67 (African-American) * .55 (28–37) = 14.6 months

1987 cohort: S = 124 * .47 (high) * .72 (male) * .79 (African-American) * .52 (18–27) = 17.2 months

While the expected average time from release to return for PMA parolees with these characteristics is 14.6 months, a counterpart non-PMA 1987 parolee has an average survival time of 17.2 months. The difference in the predicted survival time reflects the notable difference across these two cohorts in not only the level but also the timing of return.

Once again, in an attempt to estimate the impact of risk on expected survival times, we substitute the coefficient for medium risk for the coefficient for high risk, leaving all of the other coefficients unchanged. The following equations were used to generate these revised predicted values.

PMA cohort: S = 194* .65 (medium) *.64 (male) *.67 (African-American) *.55 (28–37) = 29.7 months

1987 cohort: S = 124* .66 (medium) *.72 (male) *.79 (African-American) *.52 (18–27) = 24.2 months

Changing from high risk to medium risk has a substantial impact on survival time. This change produces 15.1 and 7.0 excess number of predicted survival months for the PMA and 1987 cohorts respectively. These differences in predicted values illustrate the impact of high risk on expected survival times. As indicated, the impact of high risk on the predicted time until recidivism is more evident for the PMA cohort, which supports our finding on the compositional difference in high-risk category in that cohort.

Summary of the empirical findings

On the whole, we have found that the factors thought to predict parole outcome based on our earlier descriptive results also are related to the timing of parole failure and are consistent with the results from descriptive analyses. The salience of several factors such as assessed risk, race/ethnicity, age, prior incarceration offense, and gender are reconfirmed in predicting differences in the timing of parole failure as well as parole outcome. The effects of risk are the strongest and the most consistent across the cohorts. We also observed consistently significant effects of race/ethnicity and age. With regard to race/ethnicity, the one persistent effect is that between Anglos and minorities. When significant, the effects of age were rather strong. Unlike the violent offenders whose criminal inclination is declining after their late 20s or 30s, however, paroled property offenders aged 28–37 still showed a relatively short mean time until recidivism.

In addition, the pattern of predicted values is somewhat consistent with our preliminary conclusions established with regard to trends in survival probabilities and the hazard rates in the previous two sections. The early cohorts (1984 and 1985) have very similar predicted survival times. The 1987 cohort has a predicted time that is between eight and ten months shorter than the two early cohorts. On the other hand, the 1986 cohort has an estimated survival time only one to three months shorter than the prior two cohorts, which may be partly due to the differences in the number of significant variables in the estimation equations across cohorts (e.g., three variables for 1984 and 1986 cohorts, and four variables for the 1985 and 1987 cohorts).

In addition, we compared PMA releasees with the 1987 non-PMA group to examine the possible impact of a legislative change, i.e., accelerated release under the Prison Management Act, on the predicted length of time from release to reincarceration. The results for the PMA cohort are identical to the results for the non-PMA cohort in terms of statistically significant variables—risk assessment score, race/ethnicity, age, and gender have significant net effects on the survival time. However, there are important differences in terms of the magnitude of the effects. This finding is more evidence that indicates a possible reduction in the deterrent influence on paroled property offenders who experienced accelerated early release under PMA.

In this regard, we estimate the expected (or mean) survival time for parolees with characteristics that are associated with a high likelihood of return. While the expected average survival time for PMA parolees with these characteristics is 14.6 months, a counterpart non-PMA 1987 parolee has an average survival time of 17.2 months. The difference in the predicted survival time between these two cohorts reflects differences not only in the level but also the timing of return.

As suggested in previous studies (Ekland-Olson et al., 1991, pp. 126–128; Joo, 1993), this finding suggests that factors in each cohort which maximize the level and the timing of return to prison changed over time. This is in part due to the compositional differences between two groups, but also partly due to administrative changes in the policy of parole revocations as well as legislative changes (e.g., PMA) that may have lessened the deterrent effect of incarceration. While this analysis does not allow us to measure the relative importance of the factors, it is clear that compositional differences play an important role, as is evidenced by the substitution of the coefficient for medium risk for the coefficient for high risk. It is also noteworthy that the differences between PMA and non-PMA cohorts in the predicted survival time, as well as in the level and the timing of return, are also consistent with a reduced deterrent effect.

The analysis of survival time provides important information about the pattern of return, which allows policymakers to identify parolee characteristics that are related to the timing of reincarceration. Besides, they can use this analysis to estimate the effect of a particular individual characteristic or a certain correctional program on the survival time, controlling for the other variables.

References

Allison, Paul (1984). *Event History Analysis: Regression for Longitudinal Event Data.* Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills/London: Sage.

Blossfeld, H-P., Alfred, H. and Mayer, K. U. (1989). *Event History Analysis.* Hillsdale: Lawrence Erlbaum.

Blumstein, A., Cohen, J., Martin, S. and Tonry, M. (eds) (1983). *Research on Sentencing: The Search for Reform*, Volumes 1 and 2. Washington, DC: National Academy Press.

Blumstein, A., Cohen, J., Roth, J. and Visher, C. (eds) (1986). *Criminal Careers and Career Criminal.* Volume 1. Washington, D.C., National Academy Press.

Box-Steffensmeier, J. M. and Jones, B. S. (2004). *Event History Modeling: A Guide for Social Scientists*. Cambridge, UK: Cambridge University Press.

Chung, Ching-Fan, Schmidt, P. and Witte, Ann D. (1991). Survival analysis: A survey. *Journal of Quantitative Criminology,* 7: 59–98.

Cohen, J. (1983). Incapacitation as a strategy for crime control: Possibilities and pitfalls. In M. Tonry and N. Morris (eds), *Crime and Justice*. Chicago: University of Chicago Press.

Ekland-Olson, S., Kelly, W., Joo, Hee-Jong, Olbrich, J. and Eisenberg, M. (1991). *Justice Under Pressure: Prison Crowding, Parole Release and Recidivism in Texas.* Final Research Report prepared for the National Institute of Justice.

Farrington, D. P. and Tarling, R. (eds) (1985). *Prediction in Criminology.* Albany, NY: State University of New York Press.

Gottfredson, D. and Tonry, M. (1987). Crime and justic*e. Annual Review of Research,* 9.

Greenwood, P. (1982). *Selective Incapacitation*. Santa Monica, CA: Rand Corporation.

Greenwood, P. and Abrahamse, A. (1982). Controlling the crime rate through imprisonment. In James Q. Wilson (ed.), *Crime and Public Policy*. San Francisco: Institute for Contemporary Studies.

Haapanen, R. (1990). *Selective Incapacitation and the Serious Offender: A Longitudinal Study of Criminal Career Patterns*. New York: Springer-Verlag.

Hanushek, E. and Jackson, J. (1977). *Statistical Methods for Social Scientists*. New York: Academic Press.

Joo, Hee-Jong (1993). Parole release and recidivism: Comparative three-year survival analysis of four successive release cohorts of property offenders in Texas. University of Texas at Austin.

Joo, Hee-Jong, Ekland-Olson, S. and Kelly, W. R. (1995). Recidivism among paroled property offenders released during a period of prison reform. *Criminology*, Vol.33 (No.3): 389–410.

Klein, S. and Caggiano, M. (1986). *The Prevalence, Predictability and Policy Implications of Recidivism*. Santa Monica, CA: Rand Corporation.

Namboodiri, K. and Suchindran, C. M. (1987). *Life Table Techniques and Their Applications*. New York: Academic Press.

Petersilia, J. (2003). *When Prisoners Come Home: Parole and Prisoner Reentry*. New York: Oxford University Press.

Rossi, Peter H., Berk, R. and Lenihan, K. (1980). *Money, Work, and Crime: Experimental Evidence.* New York: Academic Press.

Schmidt, Peter and Witte, Ann D. (1987). Predicting criminal recidivism using split population survival time models. NBER Working Paper Series No. 2445. Cambridge, MA: National Bureau of Economic Research.

Schmidt, Peter and Witte, Ann D. (1988). *Predicting Recidivism Using Survival Models.* New York: Springer-Verlag.

Singer, Judith D. and Willett, John B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Sorenson, A. (1977). Estimating rates from retrospective questions. In D. Heise (ed.), *Sociological Methodology*, pp. 209–223. San Francisco: Jossey-Bass.

Tonry, M. (1987). Prediction and classification: Legal and ethical issues. In M. Tonry and N. Morris (eds), *Crime and Justice: An Annual Review of Research*. Chicago: University of Chicago Press.

Tuma, N. and Hannan, M. (1978). Approaches to the censoring problem in analysis of event histories.

In K. Schuessler (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.

Vogt, W Paul (1999). *Dictionary of Statistics and Methodology*: *A Nontechnical Guide for the Social Sciences*, 2nd edn. Thousand Oaks, CA: Sage.

von Hirsch, A. and Gottfredson, D. (1983–84). Selective incapacitation: Some queries about research design and equity. *Review of Law and Social Change,* 12: 11–51.

Yamaguchi, Kazuo (1991). Event history analysis. *Applied Social Research Methods,* Vol. 28. Newbury Park: Sage.

This page intentionally left blank

Chapter 27

Discrete-time survival analysis: predicting *whether*, and if so *when*, an event occurs

Margaret K. Keiley, Nina C. Martin, Janet Canino, Judith D. Singer and John B. Willett

1 Introduction

Researchers often ask *whether* and, if so, *when*, critical events in the life course occur. These questions are often difficult to address because of a problematic information shortfall, known as *censoring*, that often occurs when not everyone in the sample experiences the target event during the period for which the data were collected. In this chapter, we show how the methods of *discrete-time survival analysis* (aka *event history analysis* and *hazard modeling*) are ideal for studying event occurrence because they allow the even-handed incorporation of data from both the noncensored and censored cases alike. We use retrospective longitudinal data on the ages at which adolescents self-report that they had their *first experience of sexual intercourse* (the target event) to introduce fundamental statistical quantities, the *hazard* and *survival probability*. Then, we generate, specify, explain, fit, and interpret formal discrete-time hazard models of the relation between the risk of event occurrence and critical predictors, including predictors that describe the passage of time itself. Finally, we describe how researchers who pose research questions about *whether* and *when* but who choose *not* to adopt a survival-analytic framework can easily be led astray by traditional statistical analysis.

An important class of research question often posed in the social sciences asks “whether” and, if so, “when” a target event occurs. Researchers investigating the consequences of childhood traumas on later well-being, for instance, ask *whether* an individual ever experiences depression and, if so, *when* onset first occurs (Wheaton, Roszell and Hall, 1997). Other researchers ask questions about whether and when street children return to their homes (Hagan and McCarthy, 1997), whether and when college students drop out of school (DesJardins, Ahlburg and McCall, 1999), whether and when recently married couples get divorced (South, 2001) and whether and when adolescent boys (Capaldi, Crosby and Stoolmiller, 1996) or university students (Canino, 2002; 2005) experience sexual intercourse for the first time.

Familiar statistical techniques, such as multiple regression analysis and analysis of variance, and even their more sophisticated cousins, such as structural equation modeling, are ill-suited for addressing questions about the occurrence and timing of events. These usually versatile methods fail because they are unable to handle situations in which the value of the outcome—i.e., *whether* and *when* an event occurs—is unknown for some of the people under study. When event occurrence is being investigated, however, this type of information shortfall is commonplace if not inevitable. No matter how long a researcher is funded to collect data, some people in the sample may not experience the target event while they are being observed—some adults will not become depressed, some street children will not return to their parental homes, some college students will not drop out of school, some recently married couples will not divorce, some adolescents will remain virgins. Statisticians say that such cases are *censored*.

Censoring creates an important analytic dilemma that cannot be ignored. Although the investigator knows something important about the individuals with censored event times—if they do ever experience the event, they will do so *after* the observation period (the period for which the data were collected) ends—this knowledge is imprecise. If a university student does not experience sexual intercourse by age 21, for example, we would not want to conclude that he or she will *never* do so. All we can say is that by age 21, he or she was still a virgin. Yet, the need to incorporate data from individuals with censored and noncensored event times simultaneously into respectable data analyses is clear because the censored individuals are not a random subgroup of the sample. They are a special group of people—the ones who are *least likely* to experience the event, the ones who are the "longest-lived" participants in the sample. Consequently, they provide considerable information about the potential rarity of target event occurrence. Credible investigation of event occurrence requires a data-analytic method that deals evenhandedly with both the noncensored and the censored observations. Biostatisticians modeling human lifetimes to the event of death were initially stimulated to develop a class of appropriate statistical methods for analyzing such data because they faced the censoring problem constantly in medical research where, often (and thankfully), substantial numbers of their study participants did not die by the end of the observation period (Cox, 1972; Kalbfleisch and Prentice, 1980). Despite the foreboding appellations of the techniques that were thus developed – they became known variously as *survival analysis*, *event history analysis*, and *hazard modeling*—these techniques have now become invaluable to social scientists outside the medical field because they provide a sound and reasonable statistical basis for exploring the "whether" and "when" of all kinds of interesting target events in the lives of participants.

In this chapter, we provide a conceptual introduction to these survival methods, focusing specifically on the principles of discrete-time survival analysis. After distinguishing between discrete-time and continuous-time survival analysis and explaining why we encourage first-time learners to begin with the former approach, we use an example of retrospective longitudinal event-history data on the age at first sexual intercourse for a sample of college-going adolescents to introduce the fundamental building blocks of these methods. These building blocks are known as the *hazard* and *survival probabilities,* and they offer two complementary ways of describing patterns in the risk of event occurrence over time. We then introduce and specify statistical models that can be used to link these temporal patterns of risk to selected predictors, including time itself, and we comment on the types of predictors that can easily be included in these models. We then show how discrete-time hazard models can be fitted to longitudinal data and model parameters estimated, tested, and interpreted. Finally, we comment on how easily researchers can be misled if they resort to traditional data-analytic techniques for addressing these same questions, instead of adopting survival methods. Our presentation is intended to be conceptual and nontechnical–readers who are interested in learning more about these methods should consult Singer and Willett (2003) for additional guidance.

2 Measuring time and recording event occurrence

Before event occurrence can be investigated, a researcher must first record how long it takes, from some agreed-upon starting point, for each individual in a sample to experience the target event, or be censored. The researcher must thus be able to define a "beginning of time" in some meaningful and unambiguous way, must establish a suitable metric in which the passage of time can be recorded, and must irrevocably recognize the target event when it occurs. We comment on each of these three points briefly below.

Depending on the particular research project in question, investigators possess a great deal of flexibility in identifying the "beginning of time." Often, because physical birth is both handy and meaningful across a wide variety of substantive contexts, many researchers choose it as the "beginning of time" and consequently use an individual's *age* (i.e., *time since birth*) as the metric in which time is measured (see, e.g., Wheaton et al., 1997). But researchers need not restrict themselves to birth and to the metric of chronological age. One way of establishing a beginning of time in a particular study is to tie it to the occurrence of some other *precipitating event*—one that places all individuals in the population *at risk* of experiencing the forthcoming target event. When modeling how long it takes before street children return to their parental home, for example, the "beginning of time" is naturally defined as the time at which the prospective street children left their parental home for the streets for the first time (thereby making "time on the street" the metric for the ensuing survival analysis).

Once a common start time has been defined, the researcher must observe participants—either *prospectively* as time passes, or using *retrospective* reconstruction of the event history—to record *whether* and, if so, *when,* the target event occurs to each. All participants who experience the target event during the observation period are then assigned an event time equal to the time at which they actually experienced the event. Individuals who do not experience the target event during the window of the observation period are assigned *censored event times*, set equal to the time at which the observation period ended or when the individual was no longer at risk of experiencing the event, but labeled as "censored" to indicate that the target event had not occurred by that time. These censored event times, although seemingly inconclusive, tell us a great deal about the distribution of event occurrence because they establish that participants did not experience the target event at any time up to and including their time of censoring.

In some research, investigators can record event occurrence very precisely indeed. When studying the relationship between experiences of childhood adversity and subsequent death, for example, Friedman, Tucker, Schwartz and Tomlinson-Keasey (1995) used public records to determine the precise time—in years, months, and even days—when each individual who had died had actually passed away. Other researchers can record only that the target event occurred within some discrete-time *interval.* A researcher might know, for example, the *year* that a person first experienced depressive symptoms or first had sexual intercourse, the *month* when an individual began a new job, or the *grade* when a youngster transitioned from adult-supervised care to self-care. We distinguish between these two scales of measurement by calling the former *continuous-time* data and the latter *discrete-time* data.

In this chapter, we focus on statistical methods for analyzing event occurrence recorded in discrete time. We have several reasons for our emphasis. First, we have found that discrete-time methods are intuitively more comprehensible than their continuous-time cousins, facilitating initial mastery and subsequent transition to continuous-time methods. Second, we believe that discrete-time methods are highly appropriate for much of the event history data being collected naturally by social scientists because, for logistical and financial reasons, these data are often recorded only in terms of discrete intervals (see Lin, Ensel and Lai, 1997). Third, the discrete-time approach facilitates inclusion of both *time-invariant* and *time-varying* predictors, whereas inclusion of the latter is more difficult under the continuous-time approach. Thus, with discrete-time survival analysis, researchers can easily examine the impact of predictors, such as family structure and employment status, whose values fluctuate naturally over the life course. Fourth, discrete-time survival analysis explicitly fosters inspection of how patterns in the risk of event occurrence unfold over time, whereas the most popular continuous-time survival-analytic strategy ("Cox regression"; Cox, 1972) ignores the shape of the temporal risk profile entirely in favor of estimating the influence of predictors on that profile, under a restrictive assumption of "proportionality." Fifth, under the discrete-time approach, the proportionality assumption is easily assessed and "nonproportional" models specified, fitted, and interpreted. Finally, in discrete-time survival analysis, all model-fitting and parameter estimation can be conducted using standard statistical software that has been designed for standard logistic regression analysis. The researcher can thus avoid reliance on the dedicated computer software required for continuous-time survival analyses and employ good practices of sensible data analysis, as all of the investigator's usual analytic skills can be brought to bear. In our case, we used the *Statistical Analysis System*, PC-SAS, Version 9.1 (SAS Institute, 2005) to fit our hypothesized discrete-time hazard models.

3 Descriptive analysis of discrete-time survival data

The *hazard probability* and the *survival probability* are the two fundamental quantities at the center of all discrete-time survival analysis. Estimates of these probabilities provide answers to the two key questions that are usually asked of event history data: "When is the target event most likely to occur?" and "How much time passes before people are likely to experience the event?" respectively. In what follows, we introduce these two quantities and illustrate how they can be estimated and interpreted to address such questions.

In our explanation, we make use of an example of retrospective data gathered on a sample of 618 university students from a large, midwestern public university. These students were invited to earn extra classroom credit by completing a confidential online web survey (made available online from August 30–October 31, 2004) about their sexual history, level of religiosity, the quality of their romantic attachment style, and selected demographic information. The sample included 444 women (72%) and 174 men (28%). Most respondents (87%) had at least a high-school-level education. Eighty-six percent of the participants reported their race as Caucasian and the remaining 13% of respondents were African-American (5%), Hispanic (3%), Asian or Pacific Islander (3%), or other (2%). The majority of the respondents (86%) were single, 9% were engaged to be married, 4% were married, and 1% were divorced (Canino, 2005).

3.1 Person-period dataset

An important precursor to any kind of data analysis is to establish a sensible format for storing the data that you intend to analyze. To conduct discrete-time survival analyses, you must assemble your data on event occurrence in a *person-period format* in which each person contributes *one record* (row) to the dataset for *each discrete-time period* in which he/she is at risk for event occurrence. In Table 27.1, we provide an example of the person-period format using data on two participants from our "first experience of sexual intercourse" dataset. In this dataset, time is divided into discrete yearly "bins," corresponding to ages 13 through 21, none of the participants having experienced the target event at age 12 or earlier. Each participant then contributes rows to the dataset, corresponding to the years in which he or she is at risk of first experiencing sexual intercourse. For instance, notice that person #4 contributes three rows to the dataset, for ages 13 through 15, and person #151 contributes nine rows, for ages 13 through 21. Other participants contributed in a similar fashion, the number of rows that each added being determined by their sexual history, or the occurrence of censoring.

Beyond the first two columns, each subsequent column of the dataset then contains values of several classes of variables that we incorporate into subsequent analyses. These variables record important features of the problem under investigation, including: (a) the passage of time, (b) the occurrence of the target event (or the occasion of censoring), and (c) the values of important predictors that ultimately become the centerpiece of our discrete-time survival analysis. We discuss each of these three classes of variable, briefly, below.

Ultimately, we will treat participant age as a *predictor* in a forthcoming discrete-time survival analysis to investigate how participants' risks of initial sexual experience differ with age. And, as you will see, it will prove convenient at that point to represent participant age in its most general specification—as a system of dummy variables. Thus, in columns #3 through #11 of the person-period dataset, we use an alternative specification to denote participants' ages in each discrete-time period. Rather than record the value of the participant's age as a (continuous) yearly value, as in column #2, we have created a system of dichotomous time predictors, labeled *A13* through *A21*. The values of these dichotomies are set to indicate

Table 27.1 Records (rows) that a pair of adolescents (#4 and #151) contribute to the complete person-period dataset containing longitudinal discrete-time information on the occurrence and timing of first sexual intercourse as a function of self-reported attachment style and religiosity

ID	*AGE*	*A13*	*A14*	*A15*	*A16*	*A17*	*A18*	*A19*	*A20*	*A21*	*EVENT*	*AVOID*	*PREOCC*	*RELIG*
4	13	1	0	0	0	0	0	0	0	0	0	0	0	18
4	14	0	1	0	0	0	0	0	0	0	0	0	0	18
4	15	0	0	1	0	0	0	0	0	0	1	0	0	18
. . .														
151	13	1	0	0	0	0	0	0	0	0	0	1	0	12
151	14	0	1	0	0	0	0	0	0	0	0	1	0	12
151	15	0	0	1	0	0	0	0	0	0	0	1	0	12
151	16	0	0	0	1	0	0	0	0	0	0	1	0	12
151	17	0	0	0	0	1	0	0	0	0	0	1	0	12
151	18	0	0	0	0	0	1	0	0	0	0	1	0	12
151	19	0	0	0	0	0	0	1	0	0	0	1	0	12
151	20	0	0	0	0	0	0	0	1	0	0	1	0	12
151	21	0	0	0	0	0	0	0	0	1	0	1	0	12

the participant age to which each discrete-time period pertains. In the row corresponding to age 13, for instance, dummy variable *A13* takes on the value of 1 and all other time-indicators, *A14* through *A21*, are set to zero. At age 14, dummy variable *A14* is set to 1, with *A13* and *A15* through *A21* assuming the value zero, and so on. Later, we will choose to use this set of time-dummies as initial predictors in discrete-time hazard modeling to establish the age-profile of the risk of first sexual intercourse. Here, though, we simply emphasize that discrete-time survival analysis does not *require* that you specify the time predictor in this very general way as a set of dummy predictors—there would be nothing to prevent you from using participants' (linear) *AGE* as a predictor in subsequent survival analysis, for instance. However, in our empirical research, we have found that the general specification of time established here, and illustrated in Table 27.1, usually provides the most successful starting point for discrete-time survival analysis, given the typically irregular nature of any risk profile with age. For this reason, we always recommend that you establish a general specification for the time predictor up-front, when you first set up your person-period dataset.

In the twelfth column of our person-period dataset in Table 27.1, we include the all-important "event" indicator, which we have labeled *EVENT*. This variable will ultimately serve as the *outcome* variable in our discrete-time survival analyses of the relationship between risk of first sexual experience and predictors. *EVENT* is also a dichotomous variable, coded so that it takes on the value 0 at each age in which the respondent did *not* experience the target event and 1 at the single age at which the event was experienced. A key feature of the person-period dataset is that, for each participant, once the event indicator has been coded 1 (and the target event has therefore occurred), no additional records are included in the person-period dataset for that individual. An individual who experiences the event of interest—in this case, first experience of sexual intercourse—is no longer at risk of subsequent initiation, by definition, and therefore drops out of the risk set for this event. In our example, person #4 experienced sexual intercourse at the age of 15, thus *EVENT* takes on the value "0" in each of the time periods prior to age 15, but switches to value "1" in the discrete-time bin corresponding to age 15. Then, once person #4 has experienced the target event, he is no longer at risk of experiencing sexual intercourse for the first time ever again, thus he contributes no further records to the person-period dataset. By contrast, person #151 had not experienced sexual intercourse for the first time by age 21. For her, *EVENT* is coded "0" in all nine discrete-time periods in the dataset, from ages 13 through 21, and its value never switches from 0 to 1, meaning that the study ended without her experiencing the target event. She is therefore censored at age 21.

Finally, the person-period dataset also contains the values of predictors whose relationship with the risk of first sexual experience is under investigation. In our example, we have chosen to work with two important time-invariant predictors, representing the *romantic attachment style* and *religiosity* of the participants, although it would have been easy to include many others. Romantic attachment style is a categorical predictor describing three distinct states of romantic attachment, labeled "*secure*," "*avoidant*," and "*preoccupied*" (based on Brennan, Clark and Shaver's [1998] *experiences in close relationships* [ECR] measure, which assesses adult and adolescent romantic relationship attachment styles). Of the 618 adolescents in our sample, 214 were categorized with a *secure* attachment style, 201 with a *preoccupied* style, and 203 with an *avoidant* style. In our analyses, in order to include attachment as a predictor in subsequent discrete-time survival analyses, we created three dummy predictors to represent each of the three attachment styles, but included only *AVOID* and *PREOCC* in our dataset and models, omitting *SECURE* as the reference category. The values of these two predictors are listed for adolescents #4 and #151 in columns 13 and 14 of Table 27.1; because we consider attachment style time-invariant in our analyses, the values of *AVOID* and *PREOCC* are identical over all discrete-time periods, within-individual. Person #4 has a *secure* romantic attachment style, as both *AVOID* and *PREOCC* are coded zero, for him; person #151 has an *avoidant* attachment style.

Our second major predictor is the adolescent's religiosity, represented by the time-invariant continuous variable, *RELIG*, which was self-reported on a five-item scale measuring the person's degree of religious commitment with the *Duke Religion Index* (DUREL; Koenig, Patterson and Meador, 1997). In Table 27.1, notice that adolescent #4 self-reports a higher level of religiosity than adolescent #151 (values of 18 versus 12), but that the values of the variable are again time-invariant in both cases. Although we do not illustrate it here, inspection of the person-period format in Table 27.1 should easily convince you that it is a small step to the inclusion of additional time-varying predictors in the dataset, and consequently in any subsequent discrete-time hazard models. The values of time-varying predictors would simply differ from row to row across the person-period dataset, within-person.

3.2 Hazard probability

Once the discrete-time event history data have been formatted and recorded appropriately in a person-period dataset you can begin to investigate the occurrence and timing of the target event by addressing the "Whether?" and "When?" questions to which we have alluded in our introduction. However, in discrete-time survival analysis, and because of the ubiquitous presence of censoring, we do not attempt to summarize and analyze time-to-event directly. Instead, when investigating the occurrence of a target event—like the event of "first experience of sexual intercourse" in our sample of adolescents—we begin by figuring out how the "risks" of event occurrence are patterned over time. Here, for example, we investigate at what ages adolescents are at *greatest risk* of first experiencing sexual intercourse, attempting to discern whether it is during their early teens, during their late teens, or in their twenties. Determining how the "risk" of first sexual intercourse differs with adolescent age will then ultimately provide us with answers to the original research questions that we posed about the "Whether?" and "When?" of sexual initiation, as we show below.

But how can we use event-history data like these to summarize best the risk of event occurrence across age, especially if some of the participants have censored event times? We begin by introducing a fundamental statistical quantity called the *hazard probability* to represent *the risk of event occurrence in each time period*. The population hazard probability in the j^{th} discrete-time period, labeled $h(t_j)$, is defined as the *conditional* probability that a randomly selected person will experience the target event in the j^{th} time period, *given that he or she has not experienced the event in an earlier time period*. Ultimately, you will come to realize that the concept of hazard probability underpins all of discrete-time survival analysis.

The hazard probabilities describe the risk of event occurrence in each discrete-time period both as *parameters* in the *population*, or as *estimates* of those parameters in the *sample*. The estimation of hazard probability in a sample is straightforward and intuitive. In each discrete-time period, you simply identify the pool of people who still remain "at risk" of experiencing the event in that period—these are the individuals who have reached this particular time period without already experiencing the event or being censored. They are referred to as the "risk set." Then, you compute the proportion of this risk set that actually experiences the target event *in* the time period, thereby obtaining an estimate of the hazard probability for this particular discrete-time period, as follows:

$$\hat{h}(t_j) = \left(\frac{\begin{array}{c}\text{\# in sample risk set who} \\ \text{experience the target} \\ \text{event in } j^{th} \text{ time period}\end{array}}{\begin{array}{c}\text{\# in sample risk set in} \\ j^{th} \text{ time period}\end{array}} \right) \quad (1)$$

Anyone who experiences the event, or is censored, within the current period is removed from the risk set for the following time period and is therefore not included in the estimation of the hazard probability for that subsequent period. Notice how this definition of hazard probability is inherently *conditional*: it is only those individuals who have not experienced the target event, or have not been censored, in an earlier discret-time period, who can participate in the estimation of the hazard probability in a subsequent time period.

We often plot the complete set of hazard probabilities against the time period to which each refers, joining the plotted points with line segments, yielding a profile of risk with age or *hazard function*. In the left-hand plot in the top panel of Figure 27.1, we present *sample* hazard functions for our adolescents who are approaching sexual initiation. We estimated these sample hazard probabilities separately for adolescents with each of the three different romantic attachment styles—*secure*, *avoidant*, and *preoccupied*—using contingency table analysis (for a description of how to implement this approach see Singer and Willett, 2003). The obtained sample hazard functions describe the "risk" of first experiencing sexual intercourse at each of nine successive discrete ages, 13 through 21. Inspection of these functions helps pinpoint when the target event is most likely, and least likely, to occur. For instance, notice that for adolescents of all three romantic attachment styles, the risk of experiencing first sexual intercourse is relatively low at ages 13 and 14, generally increases between age 15 and 17 (with a small decrease for *preoccupied* adolescents at age 17 that may be due to sampling idiosyncrasy), and then peaks at age 18. After this age, the risk of being sexually initiated, *among those adolescents who have not yet experienced initial intercourse*, declines but, by age 21, still remains at levels greater than those experienced in early adolescence. Beyond this overall temporal profile of risk, and ignoring minor differences in shape by group, notice that across time an interesting aggregate difference in risk by attachment style occurs. *Preoccupied* adolescents appear to be consistently at the greatest risk for sexual initiation, *avoidant* adolescents appear least at risk, and *secure* adolescents enjoy a more intermediate risk. We return to these aggregate differences by attachment style later, as a way of generating formal statistical models of the hypothesized relationship between the risk of sexual initiation and predictors, like attachment style.

The "conditionality" inherent in the definition of hazard is central to discrete-time survival analysis. It ensures that all individuals—whether they are ultimately censored or experience the target event—remain in the risk set until the last time period in which they are eligible to experience the event (at which point they are either censored or they experience the target event). For example, the sample hazard probability of initial sexual intercourse at age 18 for adolescents with a *secure* attachment style is estimated, conditionally, using data on all those *secure* age 18 adolescents (124 out of the initial sample of 214) who had not yet first experienced sexual intercourse at an earlier age and who remained at risk at age 18. Of these 124 *secure* at-risk adolescents, 42 had sex for the first time at age 18, leading to an estimated sample hazard probability of 42/124, or 0.34. The conditionality of the definition of hazard probability is crucial because it ensures that the sample hazard probability deals evenhandedly with censoring—using *all* the information available

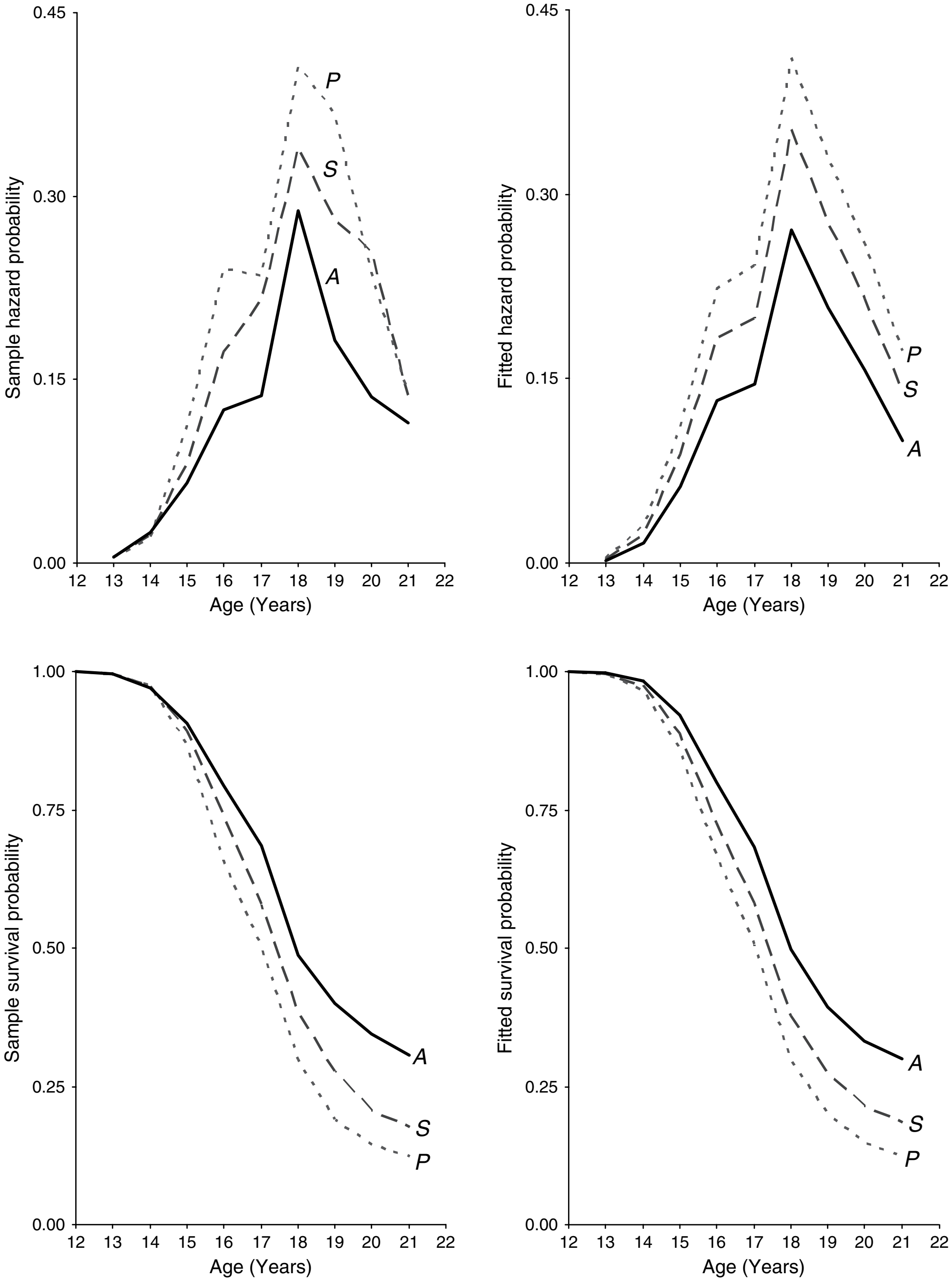


Figure 27.1 Hazard and survivor functions describing how the risk of first sexual intercourse depends upon adolescent age for 618 college students in a large mid-western university, by romantic attachment style (A = *Avoidant*; P = *Preoccupied*; S = *Secure*). The left-hand panel presents *sample* functions; the right-hand panel presents *fitted* functions from *Model 1* of Table 27.2.

in the sample event histories by including both uncensored and censored cases in the risk set in any time period but not overextending this information beyond the time when the case can legitimately contribute data to the analysis.

3.3 Survival probability

In addition to using hazard probability to explore the conditional risk of event occurrence *in each discrete time-period*, you can also cumulate the period-by-period risks to provide a picture of the overall proportion of the starting samples that "survive" through each discrete time-period—i.e., that do *not* experience the event up through the time period in question. This quantity is referred to as the *survival probability*. In any given discrete-time period, it represents the probability that a randomly selected population member will continue beyond the current time period without experiencing the target event.

We can obtain values of the survival probability easily, in each discrete time-period, by accumulating the impact of the consecutive hazard probabilities. For instance, to begin, in the population, no one has yet experienced the target event, thus the survival probability is 1.00 or 100%, by definition, at the origin of time. In the first discrete time-period in which target events can occur, then, the risk set contains all members. The hazard probability is $h(t_1)$, *however,* and therefore its complement—i.e. $\{1-h(t_1)\}$—describes the proportion of the 1st-period risk set that do not experience the target event in the period. Providing that censoring has occurred at random, then $\{1-h(t_1)\}$ of the original 100% of members must survive through the first discrete time-period, and the corresponding survival probability for the 1st discrete time-period, which we label $S(t_1)$, is simply equal to $\{1-h(t_1)\}$ *of 1.00*. We can write this as $\{1-h(t_1)\} \times 1.00$, or simply $\{1-h(t_1)\}$. Similarly, in the 2nd discrete time-period, hazard probability becomes $h(t_2)$, indicating that a fraction $\{1-h(t_2)\}$ will now survive the period. Hence, the survival probability for the 2nd period must be $\{1-h(t_2)\}$ *of the proportion who survived the 1st period*. This quantity is equal to $\{1-h(t_2)\}$ of $S(t_1)$, or $\{1-h(t_2)\}S(t_1)$. The same algorithm can be repeated in each discrete time-period, successively, such that the survival probability in any discrete time-period is simply equal to the complement of the hazard probability in that period multiplied by the survival probability in the previous period. In the population, this algorithm can be represented, as follows:

$$S(t_j) = \{1-h(t_j)\}\, S(t_{j-1}) \tag{2}$$

Sample estimates can be obtained by simply substituting the corresponding sample statistics into this formula (for a lengthier description of this accumulation method and examples, see Singer and Willett, 2003).

In parallel with our earlier usage for hazard, we use the term *survivor function* to refer to a display of the complete set of survival probabilities plotted against the time periods to which they belong. In the left-hand plot of the bottom panel of Figure 27.1, we display sample survivor functions describing the cumulative sexual initiation of adolescents with different romantic attachment styles. We obtained these estimated functions by applying the algorithm in equation (2) to the sample hazard functions in the top left plot. The sample survivor functions indicate that the proportion of adolescents who "survived"—did *not* experience sexual intercourse—through each successive time period from age 13 through 21. Notice that the curves are fairly high in the beginning of time, and close to a value of 1. Subsequently, in each of the three groups defined by attachment style, the sample survivor functions drop as time passes. At the beginning of time (age 12), all adolescents are "surviving"—none of them has had sexual intercourse—thus the sample survival probabilities are 1.00, by definition. Over time, as adolescents initiate sexual activity, the sample survivor functions drop. Because most adults do end up having sexual intercourse *at some time* in their lives, the curves do tend to decline towards a lower asymptote of zero, ending in this sample at sample survival probabilities of .18 for *secure*, .12 for *preoccupied*, and .31 for *avoidant* adolescents, by age 21. These sample proportions indicate that, by the end of their 21st year, an estimated 18% of *secure*, 12% of *preoccupied*, and 31% of *avoidant* adolescents *had not* experienced sexual intercourse. By subtraction, we therefore know that 81% of *secure*, 88% of *preoccupied*, and 69% of *avoidant* adolescents *had* experienced sexual intercourse *at some point before the end of their 21st year.* Notice that all sample survivor functions tend to have a similar shape — a monotonically nonincreasing function of age or time. The details of the shape and the rate of decline, however, can differ considerably across groups. For example, although the three sample survivor functions in Figure 27.1 have a similar monotonically declining trajectory, the sharper decline among *preoccupied* adolescents suggests that, in comparison to *secure* and *avoidant* adolescents, they are more rapidly sexually initiated between the ages of 13 and 21.

Having examined these details of the risk of sexual initiation, we can now respond specifically to the "how long" question: On "average," how long does it take before an adolescent has sexual intercourse for the first time? Such questions cannot be answered by sample averages, because of the presence of censoring, but they can be answered by the estimation of a *median lifetime* from the sample survivor function. The median lifetime is the length of time that must pass until the value of the survivor function reaches one half, or .50. In other words, it is the time by which half of the individuals in the study have experienced the target event. In our example, for *secure* adolescents at the end of age 17, the sample survivor function is just above .50. At the end of age 18 it is less than .50. We can therefore use linear interpolation to estimate that this group has a median lifetime of 17.4 years, indicating that *secure* adolescents wait, on average, until they are almost 17½ years old before initiating sexual intercourse. In the *avoidant* and *preoccupied* groups of adolescents, the respective sample estimates of median lifetime differ somewhat from the secure group and are about 17 years and almost 18 years, respectively. Be warned, though: for target events that are rare, you may not be able to estimate a median lifetime because the survivor function may not drop below its halfway point by the end of the observation period.

4 Modeling event occurrence as a function of predictors

Estimating sample hazard and survivor functions is a useful tool for exploring whether and when a group of individuals is likely to experience a target event, during the window of observation. These descriptive statistics can also be used to explore questions about differences between groups. When do children repeat a grade in school, and are maltreated children more likely than nonmaltreated children to experience this event (Rowe and Eckenrode, 1999)? When do adolescents first initiate sexual intercourse, and are adolescents with high levels of religiosity less likely to initiate sexual activity than those with low levels of religiosity? Both of these examples implicitly frame selected individual characteristics—such as child maltreatment and religiosity—as predictors of the risk profile describing the occurrence of the target event, grade repetition, and sexual initiation respectively. In fact, when we inspect the three sample hazard functions displayed in the upper left of Figure 27.1 and compare them to each other, we, too, are implicitly treating attachment style as a predictor of the risk profile of first sexual intercourse. We conclude that a relation may exist between the risk profile and attachment such that adolescents with a *preoccupied* attachment style appear to be more likely to experience sexual initiation at all ages.

"Eyeball" comparisons like these lack the credibility of formal statistical tests, however, making it difficult to account for the impact of sampling idiosyncrasy and to generalize back to the underlying population. Consequently, as in most quantitative analysis, our natural next step is to specify a formal statistical model that expresses our hypotheses about the relationship between the risk profile and predictors. If we are successful, we will be able to fit any such models to data in our person-period dataset, obtain formal estimates of parameters describing the impact of hypothesized predictors on the risk of event occurrence, conduct tests of that impact, and make inferences back to the underlying population.

But what is an appropriate form for a statistical model of the discrete-time hazard probability as a function of predictors? We can motivate our specification of the forthcoming models by examining the three sample hazard functions presented in the left-hand side of the top panel of Figure 27.1, on whose relative elevations we have already commented. Recall that we have coded, and inserted into our person-period dataset, two dichotomous predictors, *AVOID* and *PREOCC*, whose values distinguish adolescents with the *avoidant* and *preoccupied* attachment styles from the omitted category, a *secure* attachment style. Thus, when $AVOID = 1$ and $PREOCC = 0$, the adolescent has an *avoidant* attachment style; when $AVOID = 0$ and $PREOCC = 1$, he or she has a *preoccupied* attachment style; when *AVOID* and *PREOCC* are both zero, the adolescent has a *secure* attachment style.

Let us now imagine that we want to specify a statistical model that sensibly expresses the relationship between discrete-time hazard probability—the conceptual "outcome" of the analysis—and adolescent *attachment style*, represented by its two dummy predictors, *AVOID* and *PREOCC*. Ignoring minor differences in the shapes of the hazard functions for a moment, notice how attachment style appears to impact the sample hazard functions in the top left corner of Table 27.1. The sample hazard function for *preoccupied* adolescents ($AVOID = 0$ and $PREOCC = 1$) is generally at a "higher" elevation relative to the sample hazard function for *secure* adolescents ($AVOID = 0$ and $PREOCC = 0$), which in its turn seems placed "higher" than the profile for *avoidant* adolescents ($AVOID = 1$ and $PREOCC = 0$). So, conceptually at least, it appears as though the effects of dichotomous predictors *PREOCC* and *AVOID* is to "shift" the sample hazard profiles around vertically. How can we capture this behavior in a sensible statistical model? If we are to develop a reasonable statistical model for the relationship between the population hazard function and predictors, we must formalize our earlier conceptualization by specifying a model that permits variation in the values of *PREOCC* and *AVOID* to displace the hypothesized population hazard profiles vertically, in some fashion. We are heartened by the fact that this is not unlike the way that the inclusion of a dummy predictor, or a pair of dummy predictors, in an ordinary linear regression model would shift the relationship between a generic outcome Y and a continuous predictor X vertically.

The difference between a discrete-time hazard and a linear regression analysis, of course, is that the discrete-time hazard profile is a set of conditional probabilities, each bounded by 0 and 1. Statisticians modeling a bounded outcome, like a probability, as a function of predictors generally do not use a standard linear regression model to express the hypothesized relationship; instead they use a *logistic* model—in which the *logit* transform of the bounded outcome is represented as a linear function of predictors. This transformation is chosen as suitable for the analysis of outcomes that are probabilities because it acts to ensure that transformed values of the outcome are unbounded and, consequently, it ultimately prevents derivation of fitted values that fall outside the permissible range—in this case, 0 and 1. In fact, it is the implicit use of the logit transform in the analysis of outcomes that are probabilities that leads to the well-known technique of *logistic regression analysis* (Collett, 1991; Hosmer and Lemeshow, 2000). The logit-transformation of the population hazard probability in the j^{th} discrete-time can be represented as follows:

$$\text{logit } h(t_j) = \log_e\left(\frac{h(t_j)}{1-h(t_j)}\right) \tag{3}$$

Notice that, within the brackets on the right-hand side of this expression, we form a quotient from the hazard probability and its complement. This quotient is the ratio of the probability that an event *will* occur ($h(t_j)$) to the probability that it *will not* occur ($1-h(t_j)$) in the j^{th} discrete-time period (given that it had not occurred in an earlier time period). In other words, we recognize the quotient as a standard statistical commodity—the conditional *odds* that the target event will occur in this time period. In our example, for instance, it is the *odds* that an adolescent will experience sexual intercourse for the first time in the j^{th} time period, given that he or she had not experienced it earlier. We then take the natural logarithm of the conditional odds of event occurrence to obtain the logit-transformed or *log-odds* of hazard. This new quantity ranges between minus and plus infinity as hazard ranges between 0 and 1 and is therefore unbounded.

Now, we can specify the risk of event occurrence as a linear function of predictors without having to entertain the possibility that outrageous fitted values will obtain. Thus, instead of treating untransformed hazard as the raw outcome in our discrete-time hazard models, we treat logit-hazard as the new outcome and, in our example, we might specify a discrete-time hazard model for the risk of sexual initiation as a function of adolescent age and attachment style for individual i in discrete-time period j, as follows:

$$\begin{aligned} logit\ h_i(t_j) &= [\alpha_{13}A13_j + \alpha_{14}A14_j + \cdots \alpha_{21}A21_j] \\ &\quad + [\beta_1 AVOID_i + \beta_2 PREOCC_i] \end{aligned} \tag{4}$$

Ultimately, our intention is to fit this hypothesized model to the event history data that is recorded in our person-period dataset, in order to test, estimate, and interpret its parameters as a way of addressing our research questions. The model postulates that an outcome—in our case, the log-odds of hazard—is a linear function of two classes of predictors, here distinguished by brackets: (a) the adolescent's age, expressed as a system of nine time dummies, *A13* through *A21*, and (b) the substantive question predictors *AVOID* and *PREOCC*, jointly representing the adolescent's attachment style. We comment on each class briefly below.

The first class of predictors on the right-hand side of the discrete-time hazard model in (4), contained within the first set of brackets, provide the *baseline logit-hazard profile.* The slope parameters associated with each of the dichotomous time predictors—$\alpha_{13}, \alpha_{14}, \ldots, \alpha_{21}$—represent the population values of the outcome—now a logit-transformed hazard probability—in each discrete-time period for the group of *secure* adolescents (for whom $AVOID = 0$ and $PREOCC = 0$). Substituting these latter predictor values into the model for adolescents with a *secure* attachment style yields the following:

$$\begin{aligned} & logit\ h_i(t_j\,|AVOID_i = 0; PREOCC_i = 0) \\ &\quad = [\alpha_{13}A13_j + \alpha_{14}A14_j + \cdots \alpha_{21}A21_j] \end{aligned} \tag{5}$$

Working from this reduced model, it is easier to see how the remaining model parameters work to define the hypothesized population hazard function for the "baseline" *secure* group of adolescents. Recall that, in the person-period dataset, we have coded the values of the nine times dummies such that each one separately takes on the value 1 in the time period to which it refers. Thus, at age 13, predictor *A13* takes on the value 1 while the rest of the dummies are set to zero (see Table 27.1). Substituting these values into (5) reduces the hypothesized model to an even simpler form:

$$\begin{aligned} &logit\ h_i(t_1\,|AVOID_i = 0; PREOCC_i = 0) \\ &\quad = [\alpha_{13}\,(1) + \alpha_{14}\,(0) + \cdots \alpha_{21}\,(0)] = \alpha_{13} \end{aligned} \qquad (6)$$

This new discrete-time hazard model specification ensures that parameter α_{13} simply represents the value of population logit-hazard—the transformed risk of event occurrence—at age 13. Similar substitutions into (5) for each of the other discrete-time periods confirms that parameters $\alpha_{14}, \ldots, \alpha_{21}$ represent the population logit-hazard of event occurrence in the other discrete-time periods, respectively. Taken together as a group, then, the α-parameters are simply a logit-transformed population representation of the hazard function for adolescents with a *secure* attachment style, one per period, whose incarnation in the sample we have displayed as one of the three curves in the upper-left plot of Figure 27.1. If we were to fit this reduced model to data in our person-period dataset, we could estimate the α-parameters, detransform them (see below), and plot them to obtain the fitted hazard function for the *secure* group.

The simple discrete-time hazard model in (4) also contains question predictors *AVOID* and *PREOCC, however,* to represent the two facets of adolescent attachment style other than "*secure*." We can thus write down hypothesized expressions for the population logit-hazard of sexual initiation in these other two groups by substituting their respective values of the predictors and simplifying, as follows:

$$\begin{aligned} &logit\ h_i(t_j\,|AVOID_i = 0; PREOCC_i = 1) \qquad (7) \\ &\quad = [\alpha_{13}A13_j + \alpha_{14}A14_j + \cdots \alpha_{21}A21_j] + [\beta_2] \\ &logit\ h_i(t_j\,|AVOID_i = 1; PREOCC_i = 0) \qquad (8) \\ &\quad = [\alpha_{13}A13_j + \alpha_{14}A14_j + \cdots \alpha_{21}A21_j] + [\beta_1] \end{aligned}$$

Notice, now, how the additional β-parameters function in the hypothesized discrete-time hazard model. Comparing (7) and (5), for instance, you will see that our model specification permits the logit-hazard functions for the *preoccupied* and *secure* groups to have a similar temporal profile of risk (embodied in the magnitudes of the α-parameters), but that the profile for the *preoccupied* group is "shifted vertically" by β_1 in each discrete-time period. If parameter β_1 were positive, for instance, the logit-hazard profile describing the sexual initiation of *preoccupied* adolescents would retain the same shape as the baseline *secure* group, but be elevated above it by a distance β_1 (in units of logit-hazard). We have observed this behavior in the sample data (in the upper-left plot of Figure 27.1), where the (untransformed) hazard profile for the *preoccupied* group appears to be elevated above that of the *secure* group, ignoring minor variations in the shape of the risk profiles. By a similar argument, by comparing (8) and (5), we can argue that slope parameter, $\beta_{2,}$ represents the shift associated with being in the *avoidant* group (again relative to the baseline *secure* group), although inspection of the upper-left plot in Figure 27.1 suggests that the sign on this parameter may be negative once the model has been fitted to data.

4.1 Fitting the discrete-time hazard model to data and interpreting the results

The population discrete-time hazard model specified in (4) has features that seem to be mathematically sensible, given the sample plots we have examined in Figure 27.1. Our purpose now becomes the fitting of the model to our event history data in order to obtain relevant goodness-of-fit statistics, parameter estimates, and associated inferences with which to address the research questions about the "whether" and "when" of adolescent sexual initiation. While detailed justification of methods of statistical estimation is not possible here, technical work has shown that model fitting and parameter estimation can easily be conducted by using standard logistic regression analysis to regress the binary outcome, *EVENT*, on the time dummies and on the substantive predictors *in the person-period dataset* (see Singer and Willett, 2003, for a more complete explanation).[1] Parameter estimates, standard errors, and statistical inference obtained in these logistic regression analyses are then exactly those that are required by the discrete-time survival analysis. We illustrate this here, by fitting two discrete-time hazard models to our data on adolescent sexual initiation. The first model ("*Model 1*") contains the age-dummies and the *AVOID* and *PREOCC* predictors and is therefore used to investigate the main effects of adolescent age and attachment style. Our second model ("*Model 2*") adds the predictor *RELIG* in order to evaluate the marginal main effect of adolescent religiosity. The fitted discrete-time hazard models are presented in Table 27.2.

Table 27.2 Parameter estimates, approximate p-values, and standard errors from discrete-time hazard models representing the risk of first sexual intercourse in adolescence as a function of age, self-reported attachment style, and religiosity (n adolescents = 618, n events = 493)

Predictor	*Discrete-time hazard model*	
	#1	*#2*
A_{13}	−5.72 (0.71)	−4.70 (0.73)
A_{14}	−3.75 (0.28)	−2.73 (0.32)
A_{15}	−2.34 (0.16)	−1.31 (0.23)
A_{16}	−1.50 (0.13)	−0.46 (0.21)
A_{17}	−1.39 (0.14)	−0.33 (0.22)
A_{18}	−0.61 (0.13)	0.50 (0.22)
A_{19}	−0.96 (0.16)	0.16 (0.24)
A_{20}	−1.30 (0.20)	−0.15 (0.27)
A_{21}	−1.82 (0.26)	−0.64 (0.32)
AVOID	−0.38 (0.13)	−0.52 (0.13)
PREOCC	0.25 (0.12)	0.14 (0.13)
RELIG		−0.06 (0.009)
Deviance statistic	2488.41	2450.89

In Model 1 of Table 27.2, as explained above, the parameter estimates associated with each of the time-period dummy predictors, *A13* through *A21*, provide the fitted shape of the baseline logit-hazard profile for adolescents with a *secure* attachment style. We can detransform these estimates back into the world of regular hazard in order to present the age-profile of the risk of sexual initiation for this group. For example, the fitted logit-hazard of sexual initiation in the first discrete-time period—i.e.,

[1] To fit the discrete-time hazard model specified in (4), you must conduct your logistic regression analyses with the "no-intercept" option selected, no stand-alone intercept is included in the model specification. In fact, in our model specification in (4), the α-parameters function as a set of nine intercepts, one per discrete-time period.

at age 13—is estimated as −5.72 in the first row of Table 27.2, which means that:

$$\log_e\left(\frac{\hat{h}(t_1)}{1-\hat{h}(t_1)}\right) = -5.72$$

Exponentiating, cross-multiplying and rearranging terms in this expression, we obtain:

$$\left(\frac{\hat{h}(t_1)}{1-\hat{h}(t_1)}\right) = e^{-5.72}$$

$$\hat{h}(t_1) = e^{-5.72}\left(1-\hat{h}(t_1)\right)$$

$$\hat{h}(t_1) = \left(\frac{e^{-5.72}}{1+e^{-5.72}}\right) = \left(\frac{0.00328}{1+0.00328}\right)$$

$$\hat{h}(t_1) = 0.0032$$

The resulting value indicates that the fitted hazard probability of sexual initiation at age 13 is very small, with only about one-third of one percent of *secure* adolescents at risk at this age. A similar computation, however, provides the following fitted risk of sexual initiation at age 18 for members of the same group:

$$\hat{h}(t_1) = \left(\frac{e^{-0.61}}{1+e^{-0.61}}\right) = \left(\frac{0.5434}{1+0.5434}\right)$$

$$\hat{h}(t_1) = 0.35$$

Thus, among *secure* adolescents who had not yet had sexual intercourse prior to age 18, about 35% had intercourse for the first time at age 18. In a similar fashion, we can compute fitted hazard probabilities at all ages for the *secure* group and, from these fitted values, construct a fitted hazard function for the group. We have plotted this fitted risk profile as the middle curve in the upper-right plot of Figure 27.1.

To get some sense of the risk of sexual initiation for the other attachment subgroups, we can examine the parameter estimates associated with the two attachment predictors in Model 1, which estimate the shift in baseline logit-hazard from the baseline *secure* group to the *avoidant* and *preoccupied* groups, respectively. The negative parameter estimate associated with the *avoidant* group (−.38) indicates that adolescents with *avoidant* attachment styles are, across all ages, at *lower* risk for sexual initiation at every age than are *secure* adolescents. To obtain their fitted hazard profiles, you would simply subtract an amount .38 from the fitted logit-hazard values obtained for the *secure* group (i.e., the estimates associated with the α-parameters), and then detransform each back into the world of regular hazard (as we have shown above) and again plot. We have displayed this second fitted risk profile as the lower curve in the upper-right plot of Figure 27.1. The positive parameter estimate associated with the *preoccupied* group (.25) indicates that adolescents with *preoccupied* attachment styles are, across all ages, at *greater* risk of engaging in first sexual intercourse than are *secure* adolescents. Again, we can combine this estimate with the estimates associated with the α-parameters and detransform to obtain the fitted hazard probabilities for the preoccupied group. We have displayed this third fitted risk profile as the lower curve in the upper-right plot of Figure 27.1.

Comparing the right and left panels, in the upper panel of Figure 27.1, notice that the *fitted* hazard functions on the right side are far smoother than the *sample* hazard functions on the left side. This smoothness results from constraints inherently imposed in the population hazard model specified in (4), which forces the vertical separation of the logit-hazard functions of any pair of groups to be identical in each time period. Just as we do not expect a fitted regression line to touch every data point in a scatterplot, we do not necessarily expect every point on a fitted hazard function to match every sample value of hazard at the corresponding age. Indeed, in this example, further analysis not presented here reveals that the discrepancies between the sample and fitted plots presented in the upper row of Figure 27.1 are due to idiosyncratic sampling variation, exactly analogous to the differences detected between the fitted and observed values computed in regular regression analysis. From these fitted hazard functions, we have used the cumulative algorithm in (2) to obtain companion fitted survivor functions for adolescents with the three kinds of attachment style—these are presented in the lower-right panel of Figure 27.1, and from them we could obtain fitted median lifetimes if that were required.

The parameter estimates associated with the *avoidant* and *preoccupied* groups can also be interpreted by exponentiating them and interpreting their anti-logged values as *fitted odds-ratios*, as is common in regular logistic regression analysis. Anti-logging the parameter estimates, we obtain values of $e^{-.38}$ and $e^{.25}$, or 0.68 and 1.28 respectively, for *AVOID* and *PREOCC*. At every age, then, the fitted odds that an adolescent with an *avoidant* attachment style will have sexual intercourse for the first time are about two-thirds of the odds that a *secure* adolescent would do the same, given that they had not initiated sex earlier. Similarly, at every age, the fitted odds that an adolescent with a *preoccupied* attachment style will have sexual intercourse for the first time are slightly more than one-and-a-quarter times the odds that a *secure* adolescent would do the same, given that they had not initiated sex earlier. These are quite large effects, and they are both statistically significant, as you can tell by comparing the parameter estimates to their standard errors in Table 27.2, or by consulting the associated approximate p-values listed as superscripts there.

So, what have we learned by fitting this discrete-time hazard model to these person-period data? First, we can see the more clearly articulated profile of risk across time that is revealed by pooling information across all individuals in a single analysis, and we can use the statistics associated with the parameter estimates—standard errors and p-values—to make reasonable inferences about the population from which these data were sampled. Doing so here reveals a clear pattern of risk that is consistent with the one other previous study that included attachment style as a predictor of sexual initiation: on average, adolescents with a *preoccupied* attachment style are most likely to have sex at an earlier age than adolescents with either a *secure* or *avoidant* attachment style, respectively (Canino, 2002). Second, we can quantify the decreased risk of sexual initiation for *avoidant* adolescents, and the increased risk for *preoccupied* adolescents, in comparison to *secure* adolescents, by relying on the fitted odds-ratios that we have obtained, 0.68 and 1.28 respectively.

The fitting of discrete-time hazard models provides a flexible approach for investigating what affects event occurrence, and it is a method that evenhandedly incorporates data from both censored and nonuncensored individuals, as we have described. Although these models may appear unusual at first glance, they actually closely mirror the more familiar multiple linear and logistic regression models. If you know how to conduct multiple linear and logistic regression analysis, then you know how to do discrete-time survival analysis. Like their more familiar cousins, discrete-time hazard models can incorporate multiple predictors simultaneously, by simply adding them to the models. Inclusion of multiple predictors permits examination of the effect of one predictor while controlling statistically for the effects of others.

We have illustrated the inclusion of multiple predictors by adding predictor *RELIG* to obtain *Model 2* in Table 27.2. Notice that the change in the deviance statistic between *Models 1* and 2—a difference of $(2488.41-2450.89)=37.52$, for a loss of 1 degree of freedom – confirms that religiosity is a statistically significant predictor of adolescent sexual initiation, after controlling

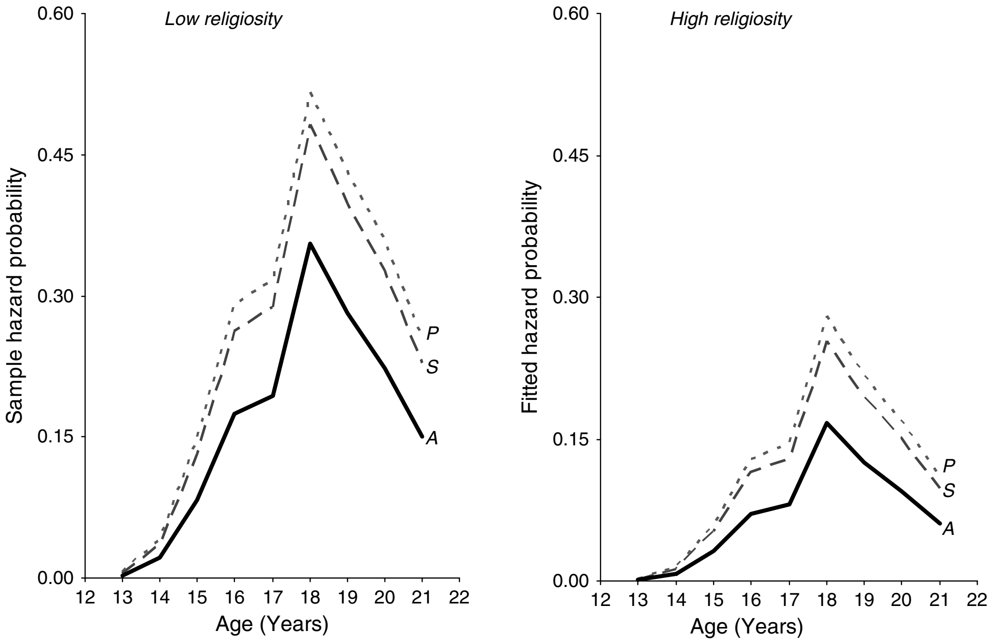


Figure 27.2 Fitted hazard functions describing how the risk of first sexual intercourse depends upon adolescent age, by attachment style (*secure*, *avoidant*, and *preoccupied*) and religiosity (low = 9.6; high = 26.4), from *Model 2* of Table 27.2.

for attachment style ($p < .001$).[2] The parameter estimate associated with this newly added predictor is negative, indicating that more religious adolescents tend to be at lower risk of sexual initiation at each age, and that the anticipated decrement in logit-hazard is .06 for a 1-unit difference in adolescent religiosity. In fact, in our sample, adolescent religiosity has a standard deviation of 5.6, thus we would expect that two adolescents whose religiosities were a standard deviation different would differ in the logit-hazard of sexual initiation by ($-.06 \times 5.6$) or -0.336. Anti-logging ($e^{-.336} = .74$), therefore, we find that the fitted odds that an adolescent who is one standard deviation more religious will be about three-quarters of the odds of a less religious adolescent, at each age. This same difference, by religiosity, is evident for adolescents of each attachment style in the fitted hazard and survivor functions of Figure 27.2, which we have obtained by the usual methods.[3] Notice the relative differences in the effect sizes due to attachment style and religiosity.

[2] Differences in deviance statistic between nested discrete-time hazard models can also be used to test general linear hypotheses concerning the joint impact of multiple predictors, as in regular linear and logistic regression analysis.

[3] In Figure 27.2, we plot fitted hazard functions for prototypical adolescents whose religiosity is 1.5 standard deviations above and below the sample mean value of 8.4, for values of 9.6 (“low”) and 26.4 (“high”).

5 Extensions of the basic discrete-time hazard model

5.1 Including time-varying predictors

You can include two very different kinds of predictors in the discrete-time hazard models that we have introduced: (a) predictors whose values are *time-invariant*, and (b) those whose values are *time-varying*. As befits their label, the former are similar to the *attachment style* and *religiosity* predictors whose effects we investigated above. They describe immutable characteristics of people, such as their sex or race, and their values are stable across the lifetime. The latter, on the other hand, describe characteristics of people whose values may fluctuate over the life course, as might an individual's self-esteem, marital status, or income. As we have noted earlier, time-varying predictors are easily incorporated into the person-period dataset, with their values potentially differing from row to row, and from there it is a short and obvious step to their inclusion as predictors in the discrete-time hazard models themselves. Singer and Willett (2003) present examples of the inclusion of time-varying predictors in discrete-time hazard models and discuss subtleties of the interpretation of their fitted effects.

The ease with which time-varying predictors can be incorporated into discrete-time hazard models offers social scientists an innovative analytic opportunity. Many important predictors of the human condition fluctuate naturally with time, including family and social structure, employment, opportunities for emotional fulfillment, and perhaps most importantly, as noted above, the occurrence and timing of other events. In traditional statistical analyses, temporal fluctuation in such predictors must often be reduced to a single constant value across time, for each person. With the advent of discrete-time hazard modeling, this is no longer the case. Researchers can examine relationships between event occurrence and predictors whose values are changing dynamically.

There are at least two reasons why we believe that the ability to include time-varying predictors in discrete-time hazard models represents an exciting analytic opportunity for researchers wanting to study the occurrence of events across the life course. First, these researchers often find themselves studying behavior across extended periods of time, sometimes encompassing more than 20, 30, or even 40 years. Although researchers studying behavior across short periods of time may reasonably be able to argue that the values of time-varying predictors are relatively stable during the study period (enabling them to make use of time-invariant indicators of these time-varying features), the tenability of this assumption surely decreases as the length of time studied increases. Second, many research questions focus on *links between the occurrences of several different events.* Researchers ask questions about whether the occurrence of one stressful event (e.g., parental divorce or death of a spouse) impacts the occurrence of another stressful event (e.g., one's own divorce or the onset of depression). Although it is possible to address such questions by comparing the trajectories of individuals who have had, and who have not had, the precipitating event *at any time during the interval covered by the observation period,* this approach requires the researcher to set aside data on all individuals who experienced the precipitating event *during* the observation period. By coding the occurrence of the precipitating event as a time-varying predictor instead, data from *all individuals* may be analyzed simultaneously (see Singer and Willett, 2003, for more on this topic).

5.2 Using alternative specifications for time

In the analysis presented here, we used a system of dummy predictors to provide a general specification for the effects of time on the risk of event occurrence. Often, however, alternative specifications of the effect of time are possible, resulting in more parsimonious and equally well-fitting models. Inspection of the fitted hazard functions displayed in the upper right of Figure 27.1, for instance, suggests that the main effect of adolescent age might be effectively represented by a cubic polynomial function which would, in this case, first rise and then fall asymmetrically. To test this specific function, we could refit either of the models in Table 27.2, replacing the complete set of age-dummies currently present by an intercept and three predictors representing the effects of linear, quadratic, and cubic age. Comparing the deviance statistics of the old and new models would then permit a formal test of whether the replacement was acceptable. Subsequent steps in the survival analysis could then proceed as described above, but now based on the more parsimonious representation of the main effects of adolescent age.

5.3 Including interactions among predictors

Just as in regular multiple linear and logistic regression analysis, you can easily incorporate two-way, and higher-order, interactions among predictors in the discrete-time hazard model. The process is identical to that used in regular regression analysis—you simply form cross-products of the focal predictors in the person-period dataset and include those cross-products as predictors in the discrete-time hazard model, along with their corresponding main effects. In our example, for instance, we could have added two columns to our person-period dataset in Table 27.1 to contain cross-products of *AVOID* and *PREOCC* with *RELIG*. Subsequent addition of the new pair of *AVOID* × *RELIG* and *PREOCC* × *RELIG* two-way interactions to *Model 2* of Table 27.2, to provide a new *Model 3*, say, would have permitted us to investigate whether the impact of religiosity on the risk of sexual initiation differed by adolescent attachment style.

5.4 Including interactions with time

When processes evolve dynamically, the *effects* of both time-invariant and time-varying predictors may fluctuate over time. A predictor whose effect is constant over time has the same impact on hazard in all time periods. A predictor whose effect varies over time has a different impact in different time periods. Both time-invariant and time-varying predictors can have time-varying effects. In the present example, for instance, we would ask whether the effect of religiosity on the risk of sexual initiation differed over time. If the effect of religiosity were time-invariant, its effect on the risk of sexual initiation would be the same regardless of the age of the adolescent. If the effect of religiosity differed over time, in contrast, the predictor could have a larger effect on the risk of sexual initiation in the early years of adolescence, for example when adolescents are still living at home, than during the later teen years, when they have already moved out of the house.

The discrete-time hazard models posited so far have not permitted a predictor's effect to vary with time; they are called *proportional-odds models*. Hazard profiles represented by such models have a special property: in every time period ("t") under consideration, the effect of the predictor on logit-hazard is identical. In (4), for example, the vertical shift in the logit-hazard profile for *avoidant* adolescents is always β_1 at all ages and for *preoccupied* adolescents is always β_2 at all ages. Consequently, the hypothesized logit-hazard profiles for adolescents with the three different attachment styles have identical *shapes* since their profiles are simply shifted versions of each other. Generally, in proportional-odds models, the entire family of logit-hazard profiles represented by all possible values of the predictors share a common shape and are mutually parallel, differing only in their relative elevations. An alternative to the *proportional odds* model which is commonly used in event history analysis is the *proportional hazards* model, in which it is the raw hazard rather than the logit-hazard that is assumed to be proportional. Such models can be calculated using generally available logistic regression software but using the complementary log-log function $\mathrm{cloglog}(h(t_j)) = \ln(-\ln(1-h(t_j)))$ in place of the logit function in equation (3). This form of the dependent variable more closely parallels the proportional hazards assumption in the Cox semiparametric event history analysis model (see, e.g., Box-Steffensmeier and Young, Chapter 26 in this volume) and some parametric event history models (see, e.g., Joo, Chapter 27 in this volume). One can assess empirically which of the two models, proportional odds or proportional hazards, best fits the data.

But is it sensible to assume that the effects of all predictors are unilaterally time-constant and that all logit-hazard profiles are proportional in practice? In reality, many predictors may not only displace the logit-hazard profile, they may also alter its shape. If the effect of a predictor varies over time, we must specify a *nonproportional model* that allows the shapes of the logit-hazard profiles to differ over time. As you will recall from your knowledge of regular multiple regression analysis, when the effect of one predictor differs by the levels of another, we say that the two predictors *interact*; in this case, we say that the predictor *interacts with time*. To add such an effect into our discrete-time hazard models, we simply include the cross-product of that predictor and time as an additional predictor in its own right.

We believe that the ability to include, and test, interactions with time represents a major analytic opportunity for empirical researchers. When studying the behavior of individuals over very long periods, it seems reasonable to hypothesize that the effects of predictors will vary as people experience different life stages. Although the effects of some predictors *will* remain constant throughout the lifetime, the effects of others may dissipate, or increase, over time. We believe that it is not hyperbole to state that interactions with time are everywhere, if only researchers took the time to seek them out. Present data analytic practice (and the widespread availability of prepackaged computer programs) permits an almost unthinking (and often untested) adoption of proportional hazards models (as in "Cox" regression), in which the effects of predictors are constrained to be constant over time. Yet we have found, in a wide variety of substantive applications including not only our own work on employment duration (Murnane, Singer and Willett, 1989), but also others' work on topics such as age at first suicide ideation (Bolger et al., 1989) and child mortality (Trussel and Hammerslough, 1983), that interactions with time seem to be the rule, rather than the exception. We have every reason to believe that once researchers start looking for interactions with time, they will arise commonly. The key is to *test* the tenability of the assumption of a time-invariant effect. We refer the interested reader to Singer and Willett (2003).

6 Is survival analysis really necessary?

In this chapter, we have introduced innovative statistical methods for investigating the occurrence and timing of target events. We hope that our presentation has encouraged you to learn more about these methods, because we believe that they offer analytic capabilities that other methods do not. However, we believe that a decision to use these methods is more than a simple preference—on the contrary, we are convinced that *failure* to use these methods, when they are required, can be a downright error! Our reason for taking this strong position has to do with the obvious failure of other statistical methods—including the simpler and more traditional methods—in addressing these same research questions. Survival-analytic methods deal evenhandedly with the presence of the censored cases, which are permitted to contribute information to the analysis up until the point at which they are censored. Traditional analytic methods, in contrast, either ignore censoring or deal with it in an *ad hoc* way, leading to problems such as negatively biased estimates of aggregate event time, or disregard for variation in risk over time. Further, because of their limitations in including predictors whose values vary over time or for permitting the effects of predictors to fluctuate over time, traditional methods also suffer when compared to survival analysis. Whereas traditional methods force researchers to build static models of dynamic processes; survival methods allow researchers to model dynamic processes dynamically. For all these reasons, we invite you to investigate the possibilities offered by survival methods.

Glossary

Censoring All individuals at risk of experiencing the target event but who do not experience it during the observation period are said to be *censored*.

Hazard probability The population hazard probability in any particular discrete-time period is the conditional probability that a randomly selected population member will experience the target event in that time period, given that he or she did not experience it in a prior time period.

Hazard function A plot of population hazard probabilities versus the corresponding discrete-time periods.

Median lifetime The population median lifetime is the length of time that must pass before the population survival probability drops below a value of *one-half*. In other words, it is the time beyond which 50% of the population has still to experience the target event. The median lifetime can be thought of as an "average" time to event.

Person-period dataset A longitudinal dataset used in the conduct of discrete-time survival analysis, in which each person contributes one record for each discrete-time period in which he or she is at risk of event occurrence.

Risk set In any discrete time-period, the risk set contains only those participants who remain eligible to experience the target event in the period. At the "beginning of time" all participants must be legitimate members of the "'risk set," by definition. When a participant experiences the target event, or disappears because of censoring, he or she is no longer considered a member of the "risk set."

Survival analysis A statistical method for addressing research questions that ask *whether*, and if so *when*, a target event occurs. Discrete-time survival analysis involves the modeling of the population hazard probability as a function of predictors.

Survival probability The population survivor probability in any discrete time-period is the probability that a randomly selected population member will "survive" *beyond* the current time period without experiencing the target event. In other words, it is the probability that he or she will *not* experience the event during the current, or any earlier, time period.

Survivor function A plot of population survival probabilities versus the corresponding discrete time-periods.

Target event (outcome) The uniquely defined event whose occurrence and timing the researcher is investigating.

References

Allison, P. D. (1984). *Event History Analysis: Regression for Longitudinal Event Data*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series Number 05-046. Beverly Hills, CA: Sage.

Bolger, N., Downey, G., Walker, E. and Steininger, P. (1989). The onset of suicide ideation in childhood and adolescence. *Journal of Youth and Adolescence*, 18: 175–189.

Brennan, K., Clark, C. and Shaver, P. (1998). Self-report measures of adult attachment: An integrative overview. In J. A. Simpson and W. S. Rholes (eds), *Attachment Theory and Close Relationships*. New York: Guilford.

Canino, J. (2002). The relationship between age at first sexual intercourse, number of sexual partners, duration of first sexual relationship and attachment style. Unpublished master's thesis, Purdue University, West Lafayette, Indiana.

Canino, J. (2005). Using survival analysis to explore the relationship among attachment theory, religiosity, and sexual initiation. Unpublished doctoral thesis. Purdue University, West Lafayette, Indiana.

Capaldi, D. M., Crosby, L. and Stoolmiller, M. (1996). Predicting the timing of first sexual intercourse for at-risk adolescent males. *Child Development*, 67: 344–359.

Collett, D. (1991) *Modeling Binary Data*. London, UK: Chapman & Hall.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34: 187–202.

DesJardins, S. L., Ahlburg, D. A. and McCall, B. P. (1999), An event history model of student departure. *Economics of Education Review*, 18: 375–390.

Friedman, H. S., Tucker, J. S., Schwartz, J. E. and Tomlinson-Keasey, C. (1995). Psychosocial and behavioral predictors of longevity: The aging and death of the "Termites." *American Psychologist*, 50: 69–78.

Ginexi, E. M., Howe, G. W. and Caplan, R. D. (2000). Depression and control beliefs in relation to reemployment: What are the directions of effect? *Journal of Occupational Health Psychology*, 5: 323–336.

Hagan, J. and McCarthy, B. (1997). Intergenerational sanction sequences and trajectories of street-crime amplification. In I. H. Gotlib and B. Wheaton (eds), *Stress and Adversity over the Life Course: Trajectories and Turning Points*. New York: Cambridge University Press.

Hosmer, D. W., Jr. and Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Koenig, H., Meador, K. and Parkerson, G. (1997). Religion index for psychiatric research: A 5-item measure for use in health outcome studies. *American Journal of Psychiatry*, 154: 885–886.

Lin, N., Ensel, W. M. and Lai, W. G. (1997). Construction and use of the life history calendar: Reliability and validity of recall data. In I. H. Gotlib and B. Wheaton (eds), *Stress and Adversity over the Life Course: Trajectories and Turning Points*. New York: Cambridge University Press.

Murnane, R. J., Singer, J. D. and Willett, J. B. (1989). The influences of salaries and "opportunity costs" on teachers' career choices: Evidence from North Carolina. *Harvard Educational Review*, 59: 325–346.

Roderick, M. (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal*, 31(4): 729–759.

Rowe, E. and Eckenrode, J. (1999). The timing of academic difficulties among maltreated and nonmaltreated children. *Child Abuse and Neglect*, 23(8): 813–832.

SAS Institute (2005). *Statistical Analysis System*. Available at http://www.sas.com.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Sorenson, S. G., Rutter, C. M. and Aneshensel, C. S. (1991). Depression in the community: An investigation into age of onset. *Journal of Consulting and Clinical Psychology*, 57: 420–424.

South, S. J. (2001). Time-dependent effects of wives' employment on marital dissolution. *American Sociological Review*, 66: 226–245.

Trussel, J. and Hammerslough, C. (1983). A hazards-model analysis of the covariates of infant and child mortality in Sri Lanka, *Demography*, 20: 1–26.

Wheaton, B., Roszell, P. and Hall, K. (1997). The impact of twenty childhood and adult traumatic stressors on the risk of psychiatric disorder. In I. H. Gotlib and B. Wheaton (eds), *Stress and Adversity over the Life Course: Trajectories and Turning Points*. New York: Cambridge University Press.

This page intentionally left blank

Part VI

Panel Analysis, Structural Equation Models, and Multilevel Models

This page intentionally left blank

Chapter 28

Generalized estimating equations for longitudinal panel analysis

Joseph M. Hilbe and James W. Hardin

Correlated datasets arise from repeated measures studies where multiple observations are collected from a specific sampling unit (a specific bank's accounts receivable status over time), or from clustered data where observations are grouped based on a shared characteristic (banks in a specific zip code). When measurements are collected over time, the term "longitudinal" or "panel data" is often preferred. The generalized linear models framework for independent data is extended to correlated data via the generalized estimating equations framework. We discuss the estimation of model parameters and associated variances via generalized estimating equation methodology.

1 Introduction

Parametric model construction specifies the systematic and random components of variation. Maximum likelihood models rely on the validity of these specified components, and then model construction proceeds from the (components of variation) specification to a likelihood and thereupon to an estimating equation. The estimating equation for maximum likelihood estimation is obtained by equating the derivative of the log-likelihood (with respect to each of the parameters of interest) to zero, and solving. Point estimates of unknown parameters are thus obtained by solving the estimating equation.

2 Generalized linear models

The theory and an algorithm appropriate for obtaining maximum likelihood estimates where the response follows a distribution in the single parameter exponential family was introduced in Nelder and Wedderburn (1972). This reference introduced the term generalized linear models (GLMs) to refer to a class of models which could be analyzed by a single algorithm. The theoretical justification of and the practical application of GLMs have since been described in many articles and books; McCullagh and Nelder (1989) is the classic reference.

GLMs encompass a wide range of commonly used models such as linear regression for continuous outcomes, logistic regression for binary outcomes, and Poisson regression for count data outcomes. The specification of a particular GLM requires a link function that characterizes the relationship of the mean response to a vector of covariates. In addition, a GLM requires specification of a variance function that relates the variance of the outcomes as a function of the mean. The ability to specify the variance as a function of the mean times a scalar constant of proportionality is a restriction of the class of models.

The derivation of the iteratively reweighted least squares (IRLS) algorithm appropriate for fitting GLMs begins with the general likelihood specification for the single parameter exponential family of distributions. Within an iterative algorithm, an updated estimate of the coefficient vector may be obtained via weighted ordinary least squares. The estimation is then iterated to convergence; e.g., until the change in the estimated coefficient vector is smaller than some specified tolerance.

For any response that follows a member of the single parameter exponential family of distributions,

$$f(y) = \exp\{[y\theta - b(\theta)]/\phi + c(y, \phi)\}$$

where θ is the canonical parameter and ϕ is a proportionality constant, and we can obtain maximum likelihood estimates of the $p \times 1$ regression coefficient vector $\boldsymbol{\beta}$ by solving the estimating equation given by

$$\Psi(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Psi_{\mathrm{i}} = \sum_{i=1}^{n} \mathbf{X}_i{}^{\mathrm{T}}(y_i - \mu_i)/[\phi V(\mu_i)][\partial\mu_i/\partial\eta_i]$$
$$= \mathbf{0}_{(p\times 1)}$$

In the estimation equation, $\mathbf{X}_i$ is the ith row of an $n \times p$ matrix of covariates $\mathbf{X}$, $\mu_i = g(\mathbf{x}_i\boldsymbol{\beta})$ represents the expected outcome $\mathrm{E}(y) = b'(\boldsymbol{\theta})$ in terms of a transformation of the linear predictor $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ via a monotonic (invertible) link function $g()$, and the variance $\mathrm{V}(\mu_i)$ is a function of the expected value proportional to the variance of the outcome $\mathrm{V}(y_i) = \phi\ \mathrm{V}(\mu_i)$. The estimating equation is sometimes called the score equation since it equates the score vector $\Psi(\boldsymbol{\beta})$ to zero.

Those involved in modeling GLMs are free to specify a link function as well as a variance function. If the link-variance pair of functions coincides with those functions from a single member of the exponential family of distributions, the resulting estimates are equivalent to maximum likelihood estimates. However, modelers are not limited to such choices. When selection of variance and link functions do not coincide to a particular exponential family member distribution, the estimating equation is said to imply the existence of a quasilikelihood; the resulting estimates are referred to as maximum quasilikelihood estimates.

The link function that equates the canonical parameter θ with the linear predictor $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ is called the canonical link. One advantage to the interpretation of results given the selection of the canonical link is that the estimating equation simplifies to

$$\Psi(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Psi_{\mathrm{i}} = \sum_{i=1}^{n} \mathbf{X}_i{}^{\mathrm{T}}(y_i - \mu_i)/\phi = \mathbf{0}_{(p\times 1)}$$

enforcing equivalence of the mean of the fitted and observed outcomes. A second advantage of the canonical link over other link functions is that the expected Hessian matrix is equal to the observed Hessian matrix.

3 The independence model

A basic individual-level model is written in terms of the n individual observations y_i for $i = 1, \ldots, n$. When observations are clustered, due to repeated observations on the sampling unit or because the observations are grouped by identification through a cluster identifier variable, the model may be written in terms of the observations y_{it} for the clusters $i = 1, \ldots, n$ and the within-cluster repeated, or related, observations $t = 1, \ldots, n_i$. The total number of observations is then $N = \Sigma_i n_i$. The clusters may also be referred to as panels, subjects, or groups. In this presentation, the clusters i are independent, but the within-clusters observations it may be correlated. An independence model, however, assumes that the within-cluster observations are not correlated.

The independence model is a special case of more sophisticated correlated data approaches (such as GEE). This model assumes that there is no correlation of the observations within clusters. Therefore, the model specification may be in terms of the individual observations y_{it}. Although the independence model assumes that the repeated measures are independent, the model still provides consistent estimators in the presence of correlated data. Of course, this consistency is paid for through inefficiency, though the efficiency loss is not always large as investigated by Glonek and McCullagh (1995). As such, this much-simplified model remains an attractive alternative because of its computational simplicity as well as its easy interpretation. The independence model also serves as a reference model in the derivation of diagnostics for more sophisticated models for clustered data (such as GEE models).

The validity of the (naive) model-based variance estimators depends on the correct specification of the variance; in turn this depends on the correct specification of the working correlation model. A formal justification for an alternative estimator known as the sandwich variance estimator is given in Huber (1967).

Analysts can use the independence model to obtain point estimates along with standard errors based on the modified sandwich variance estimator to ensure that inference is robust to any type of within-cluster correlation. While the inference regarding marginal effects is valid (assuming that the model for the mean is correctly specified), the estimator from the independence model is not efficient when the data are correlated.

It should be noted that assuming independence is not always conservative; the model-based (naive) variance estimates based on the observed or expected Hessian matrix are not always smaller than those of the modified sandwich variance estimator. Since the sandwich variance estimator is sometimes called the robust variance estimator, this result may seem counterintuitive. However, it is easily seen by assuming negative within-cluster correlation leading to clusters with both positive and negative residuals. The cluster-wise sums of those residuals will be small and the resulting modified sandwich variance estimator will yield smaller standard errors than the model-based Hessian variance estimators.

Other obvious approaches to analysis of the nested structure assumed for the data include fixed-effects and random-effects models. Fixed-effects models incorporate a fixed increment to the model for each group, while random-effects models assume that the incremental effects from the groups are perturbations from a common random distribution; in such a model the parameters (variance components) of the assumed random-effects distribution are estimated rather than the effects. In the example at the end of this entry, we consider two different distributions for random effects in a Poisson model.

4 Subject-specific (SS) versus population-averaged (PA) models

There are two main approaches to dealing with correlation in repeated or longitudinal data. One approach focuses on the marginal effects averaged across the individuals (population-averaged approach), and the second approach focuses on the effects for given values of the random effects by fitting parameters of the assumed random-effects distribution (subject-specific approach).

The population-averaged approach models the average response for observations sharing the same covariates (across all of the clusters or subjects), while the subject-specific approach explicitly models the source of heterogeneity so that the fitted regression coefficients have an interpretation in terms of the individuals.

The most commonly described GEE model was introduced in Liang and Zeger (1986). This is a population-averaged approach. While it is possible to derive subject-specific GEE models, such models are not currently supported in commercial software packages and so do not appear nearly as often in the literature.

The basic idea behind this approach is illustrated as follows. We initially consider the estimating equation for GLMs. The estimating equation, in matrix form, for the exponential family of distributions can be expressed as

$$\begin{aligned}\Psi(\boldsymbol{\beta}) &\\ = \sum_{i=1}^{n} \Psi_{\mathrm{i}} &= \sum_{i=1}^{n} \mathbf{X}_i{}^{\mathrm{T}} \mathrm{D}[\partial\mu_i/\partial\eta_{\mathrm{i}}]\mathbf{V}^{-1}(\mu_i)(y-\mu_i)/\phi \\ = \sum_{i=1}^{n} \mathbf{X}_i{}^{\mathrm{T}} &\mathrm{D}[\partial\mu_i/\partial\eta_{\mathrm{i}}]\mathbf{V}^{-1/2}(\mu_i) I_{(n\times n)} \\ &\times \mathbf{V}^{-1/2}(\mu_i)(y_i-\mu_i)/\phi = \mathbf{0}_{(p\times 1)}\end{aligned}$$

Assuming independence, $\mathbf{V}^{-1}(\mu_i)$ is clearly an $n_i \times n_i$ diagonal matrix which can be factored with an identity matrix in the center playing the role of the correlation of observations within a given group or cluster. This corresponds to the independence model we have previously discussed.

The genesis of the original population-averaged generalized estimating equations is to replace the identity matrix with a parameterized working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$. To address correlated data, the working correlation matrix imposes structural constraints. In this way, the independence model is a special case of the GEE specifications where $\mathbf{R}(\boldsymbol{\alpha})$ is an identity matrix.

Formally, Liang and Zeger introduce a second estimating equation for the structural parameters of the working correlation matrix. The authors then establish the properties of the estimators resulting from the solution of these estimating equations. The GEE moniker was applied because the model is derived through a generalization of the GLM estimating equation; the second order variance components are introduced directly into the estimating equation rather than appearing in consideration of a multivariate likelihood. There are several software packages that support estimation of these models. These packages include R, SAS, S-PLUS, Stata, LIMDEP, and SUDAAN. R and S-PLUS users can easily find user-written software tools for fitting GEE models, while such support is included in the other packages.

5 Estimating the working correlation matrix

One should carefully consider the parameterization of the working correlation matrix since including the correct parameterization leads to more efficient estimates. We want to carefully consider this choice even if we employ the modified sandwich variance estimator in the calculation of standard errors and confidence intervals for the regression parameters. While the use of the modified sandwich variance estimator assures robustness in the case of misspecification of the working correlation matrix, the advantage of more efficient point estimates is still worth this effort. There is no controversy as to the fact that the GEE estimates are consistent, but there is some controversy with regard to their efficiency. This concern centers on how well the correlation parameters can be estimated.

Typically, a careful analyst chooses some small number of candidate parameterizations. Pan (2001) also discusses the quasilikelihood information criterion (QIC) measures for choosing between candidate parameterizations. This criterion measure is similar to the well-known Akaike information criterion (AIC).

The most common choices for the working correlation **R** matrix are given by structural constraints, parameterizing the elements of the matrix as:

Table 28.1 Common correlation structures. Values are given for u ≠ v; $R_{uu} = 1$

Independent	$R_{uv} = 0$
Exchangeable	$R_{uv} = \alpha$
Autocorrelated – AR(1)	$R_{uv} = \alpha^{\lvert u-v \rvert}$
Stationary(k)	$R_{uv} = \alpha_{\lvert u-v \rvert}$ if $\lvert u-v \rvert \le k$ 0 otherwise
Nonstationary(k)	$R_{uv} = \alpha_{(u,v)}$ if $\lvert u-v \rvert \le k$ 0 otherwise
Unstructured	$R_{uv} = \alpha_{(u,v)}$

The independence model admits no extra parameters and the resulting model is equivalent to a generalized linear model specification. The exchangeable correlation parameterization admits one extra parameter. The most general approach is to consider the unstructured (only imposing symmetry) working correlation parameterization which admits $M(M-1)/2 - M$ extra parameters where $M = \max_i\{n_i\}$. The *exchangeable* correlation specification, the most commonly used correlation structure for GEEs, is also known as *equal correlation*, *common correlation*, and *compound symmetry*.

The elements of the working correlation matrix are estimated using the Pearson residual, calculated following each iteration of model fit. Estimation alternates between estimating the regression parameters $\boldsymbol{\beta}$, assuming that the initial estimates of $\boldsymbol{\alpha}$ are true, and then obtaining residuals to update the estimate of $\boldsymbol{\alpha}$, and then using estimates of $\boldsymbol{\alpha}$ to calculate updated parameter estimates, and so forth until convergence. GEE algorithms are built around a GLM basis, with a subroutine called to update values of $\boldsymbol{\alpha}$. Estimation of GEE models using other correlation structures use a similar methodology; only the properties of each correlation structure differs.

A schematic for representing how the foremost correlation structures appear is found below. Discussion on how the elements in each matrix are to be interpreted in terms of model fit can be found in Twisk (2003) and Hardin and Hilbe (2003).

INDEPENDENT

-			
0	-		
0	0	-	
0	0	0	-

EXCHANGEABLE

-			
p	-		
p	p	-	
p	p	p	-

M-DEPENDENT

-			
p1	-		
p2	p1	-	
0	p2	p1	-

AUTOREGRESSIVE

-			
p1	-		
p2	p1	-	
p3	p2	p1	-

UNSTRUCTURED

-			
p1	-		
p2	p4	-	
p3	p5	p6	-

5.1 Example

To highlight the interpretation of GEE analyses and point out the alternate models, we focus on a simple example.

These data are from a panel study on Progabide drug treatment of epilepsy. Baseline measures of the number of seizures in an eight-week period were collected and recorded as *base* for 59 different patients. Four follow-up two-week periods also counted the number of seizures; these were recorded as *s1*, *s2*, *s3*, and *s4*. The numbers recorded in the *base* variable were divided

Table 28.2 Data from Progabide study on epilepsy (59 patients over 5 weeks)

id	*age*	*trt*	*base*	*s1*	*s2*	*s3*	*s4*
1	31	0	11	5	3	3	3
2	30	0	11	3	5	3	3
3	25	0	6	2	4	0	5
4	36	0	8	4	4	1	4
5	22	0	66	7	18	9	21
6	29	0	27	5	2	8	7
7	31	0	12	6	4	0	2
8	42	0	52	40	20	23	12
9	37	0	23	5	6	6	5
10	28	0	10	14	13	6	0
11	36	0	51	26	12	6	22
12	24	0	33	12	6	8	5
13	23	0	18	4	4	6	2
14	36	0	42	7	9	12	14
15	26	0	87	16	24	10	9
16	26	0	50	11	0	0	5
17	28	0	18	0	0	3	3
18	31	0	111	37	29	28	29
19	32	0	18	3	5	2	5
20	21	0	20	3	0	6	7
21	29	0	12	3	4	3	4
22	21	0	9	3	4	3	4
23	32	0	17	2	3	3	5
24	25	0	28	8	12	2	8
25	30	0	55	18	24	76	25
26	40	0	9	2	1	2	1
27	19	0	1-	3	1	4	2
28	22	0	47	13	15	13	12
29	18	1	76	11	14	9	8
30	32	1	38	8	7	9	4
31	20	1	19	0	4	3	0
32	20	1	10	3	6	1	3
33	18	1	19	2	6	7	4
34	24	1	24	4	3	1	3
35	30	1	31	22	17	19	16
36	35	1	14	5	4	7	4
37	57	1	11	2	4	0	4
38	20	1	67	3	7	7	7
39	22	1	41	4	18	2	5

Table 28.2 (Continued)

id	*age*	*trt*	*base*	*s1*	*s2*	*s3*	*s4*
40	28	1	7	2	1	1	0
41	23	1	22	0	2	4	0
42	40	1	13	5	4	0	3
43	43	1	46	11	14	25	15
44	21	1	36	10	5	3	8
45	35	1	38	19	7	6	7
46	25	1	7	1	1	2	4
47	26	1	36	6	10	8	8
48	25	1	11	2	1	0	0
49	22	1	151	102	65	72	63
50	32	1	22	4	3	2	4
51	25	1	42	8	6	5	7
52	35	1	32	1	3	1	5
53	21	1	56	18	11	28	13
54	41	1	24	6	3	4	0
55	32	1	16	3	5	4	3
56	26	1	22	1	23	19	8
57	21	1	25	2	3	0	1
58	36	1	13	0	0	0	0
59	37	1	12	1	4	3	2

by four in our analyses to put this array of observations on the same scale as the follow-up counts. The *age* variable records the patient's age in years, and the *trt* variable indicates whether the patient received the Progabide treatment (value recorded as one) or was part of the control group (value recorded as zero).

An obvious approach to analyzing the data is to hypothesize a Poisson model for the number of seizures. Since we have repeated measures, there are many competing models. In our illustrations of these alternative models, we utilize the baseline measure as a covariate along with the *time* and *age* variables.

Table 28.3 contains the results of several analyses. For each covariate, we list the estimated incidence rate ratio (exponentiated Poisson coefficient). Following the incidence rate ratio estimates, we list the classical (for all models) and sandwich-based (for all but the gamma distributed random-effects model) estimated standard errors.

We re-emphasize that the independence model coupled with standard errors based on the modified sandwich variance estimator is a valid approach to modeling these data. The weakness of the approach is that the estimators will not be as efficient as a model including the true underlying within-cluster correlation structure. Another standard approach to modeling this type of repeated measures would be to hypothesize that the correlations are due to individual-specific random intercepts. These random effects (one could also hypothesize fixed effects) would lead to alternate models for the data.

Results from two different random-effects models are included in the table. We could hypothesize that the correlation follows an autoregressive process since the data are collected over time. However, this is not always the best choice; in the present experiment we would have to believe that the hypothesized correlation structure applies to both the treated and untreated groups.

Table 28.3 Estimated incidence rate ratios and standard errors for various Poisson models

Model	*time*	*trt*	*age*	*baseline*
Independence	0.944 (0.019,0.033)	0.832 (0.039,0.143)	1.019 (0.003,0.010)	1.095 (0.002,0.006)
Gamma RE	0.944 (0.019)	0.810 (0.124)	1.013 (0.011)	1.116 (0.015)
Gaussian RE	0.944 (0.019,0.033)	0.760 (0.117,0.117)	1.011 (0.011,0.009)	1.115 (0.012,0.011)
GEE(exch)	0.939 (0.019,0.019)	0.834 (0.058,0.141)	1.019 (0.005,0.010)	1.095 (0.003,0.006)
GEE(ar 1)	0.939 (0.019,0.019)	0.818 (0.054,0.054)	1.021 (0.005,0.003)	1.097 (0.003,0.003)
GEE(unst)	0.951 (0.017,0.041)	0.832 (0.055,0.108)	1.019 (0.005,0.009)	1.095 (0.003,0.005)

The QIC values (described in Pan, 2001) for the independence, exchangeable, ar1, and unstructured correlation structures are respectively given by −5826.23, −5826.25, −5832.20, and −5847.91. This criterion measure indicates a preference for the unstructured model over the autoregressive model. The fitted correlation matrices for these models (printing only the bottom half of the symmetric matrices) are given by

1.00			
0.51	1.00		
0.26	0.51	1.00	
0.13	0.26	0.51	1.00

1.00			
0.25	1.00		
0.42	0.68	1.00	
0.22	0.28	0.58	1.00

Note that if the exchangeable correlation structure were used, the correlation matrix would show a single value for all subdiagonal cells. A full discussion of the various correlation structures and how each are to be evaluated can be found in Hardin and Hilbe (2003).

References

Glonek, G.F.V. and McCullagh, R. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, 57: 533–546.

Hardin, J.W. and Hilbe, J.M. (2003). *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC Press.

Hilbe, J.M. (2007). *Negative Binomial Regression*. Cambridge, UK: Cambridge University Press.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings for the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, 221–223. Berkeley, CA: University of California Press.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73: 13–22.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman & Hall.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3): 370–384.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57: 120–125.

Twisk, J.W.R. (2003). *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge, UK: Cambridge University Press.

Chapter 29

Linear panel analysis

Steven E. Finkel

1 Introduction

Linear panel analysis refers to the statistical models and methods appropriate for the analysis of continuous or quantitative outcomes with data collected on multiple units (i.e., individuals, schools, or countries) at more than one point in time. This means that linear panel analysis is concerned with the various ways of analyzing change in continuous variables, in describing different patterns of change for different units, in modeling why some units change more than others and what variables are responsible for these differences. These methods may be distinguished from other kinds of longitudinal analyses, such as loglinear, transition or Markov models for analyzing over-time change in categorical outcomes, event history or duration models for analyzing temporal processes leading to the occurrence of specific events, and time-series methods for the analysis of change in continuous outcomes for a single unit over a relatively long period of time. Hence, what characterizes linear panel analysis is a focus on *continuous outcomes* for *multiple units* at *multiple points*.

It is conventional within this general rubric to make one further distinction. In some panel datasets, *time* is dominant, i.e., relatively few units have been observed for relatively long periods of time. In other data, "*N*" is dominant, i.e., many units have been observed for relatively few points in time. Although the two kinds of data have the same formal structure, time-dominant data, sometimes referred to as "time-series cross-sectional data," is typically analyzed with statistical methods rooted in the time-series econometric tradition (see Beck and Katz, 1995; Greene, 2003; see also Worrall, Chapter 15 in this volume). This chapter is concerned with the statistical methods used for the analysis of "N-dominant" panel data, typically with observations on hundreds or possibly thousands of units observed at two or more "waves," in time. Examples of "panel data" are the National Election Studies (NES) panels that track thousands of the same respondents across multiple presidential and congressional elections in 1956–58–60, 1972–74–76, and 2000–2002–2004, the multiwave US Panel Study of Income Dynamics (PSID), and the German Socio-Economic Panel (SOEP) covering some twenty-one waves of observation since 1986.[1]

There are several important motivations for analyzing panel versus cross-sectional data. Consider the hypothesis that economic performance contributes to the consolidation or stability of democratic regimes. This hypothesis

[1] Available at www.umich.edu/~nes/, psidonline.isr.umich.edu/, and www.diw.de/english/sop/ respectively.

could be tested with cross-sectional data by predicting some sample of countries' level of democracy at a given point in time with relevant economic indicators, perhaps including additional variables related to the social or political characteristics of the countries as statistical controls. Statistically significant effects of the economic variables would then be taken as supporting the hypothesis; if the researcher was confident enough in the specification of the model or otherwise sufficiently bold, he or she might even claim that economic performance had a *causal* effect on countries' level of democracy.

Yet serious obstacles exist for successful causal inference in the cross-sectional context, many of which can be dealt with more easily through the analysis of panel data. First, cross-sectional data contains no direct measure of changes in Y nor changes in X—it is implicitly assumed in such designs that by comparing units with higher and lower values on the independent and dependent variables simulates "changes" in X (economic performance) and their impact on changes in Y (democracy). But what has been conducted is simply a fixed comparison of countries or units at a given point in time, which says little directly about what happens when individuals or units change. Far better from the point of view of testing theories of social, psychological, or political change by directly observing change over time, and of course that is what longitudinal or panel data provides at its very foundation.

But the problems of causal inference in cross-sectional designs go beyond the lack of direct observation of change. First, panel data offer decided advantages in dealing with the problem of *spurious* relationships between variables, such that some outside variable *Z* not considered by the researcher is actually responsible for the observed relationship between X and Y. Of course, theoretically-relevant variables that are observed in a given dataset should always be incorporated into any statistical model to attempt to avoid this kind of bias. But with panel data, the researcher has the ability under some conditions to control for *unmeasured* variables that may be confounding the observed relationship between X and Y. In the economic performance and democracy example, variables such as a country's political culture or the degree of "entreprenuerialism" in the population may cause both economic and democratic outcomes, and to the extent that these variables are unobserved in the typical cross-national dataset, the estimated effect of economic performance on democracy will suffer from omitted variable bias. Panel data are no panacea for this problem but they offer the researcher far greater options for incorporating "unobserved heterogeneity" between units into statistical models, and controlling their potentially damaging effects, than is possible in cross-sectional analyses.

A second obstacle in cross-sectional data is that because X and Y are measured at the same time, it is difficult to determine which one of the variables can be presumed to "cause" the other. Does economic performance lead to changes in a country's level of democracy, or does the country's level of democracy lead to higher levels of economic performance (or both)? With panel data, the researcher can track the impact of changes in X, or the level of X, at *earlier* points in time with *later* values of Y, and estimates of the effects of earlier values of Y on subsequent values of X can similarly be obtained. The ability to estimate dynamic models of reciprocal causality between X and Y by exploiting intertemporal change in both variables is one of the important advantages of longitudinal data analysis in general, and the methods used for estimating these kinds of models for continuous outcomes form a large part of the statistical toolkit for linear panel analysis.

Third, it is also the case that successful causal inference depends on the accurate measurement of the variables in any statistical model.

As is well known, random measurement error in the independent variable of a bivariate model will attenuate the estimated effect of X on Y, with the direction of bias in multivariate models being completely indeterminate (Bollen, 1989; Wheaton, et al., 1977). While panel data are certainly not immune to these measurement problems, the longitudinal information contained in such data offer the researcher greater ability to model measurement error in the variables than is typically the case with cross-sectional data. As is the case with models of reciprocal causality, measurement error modeling can proceed with fewer potentially unrealistic and restrictive assumptions in panel analyses, thus strengthening confidence in the estimated causal linkages between variables.

Currently, there are several general methodological approaches or frameworks within which researchers conduct linear panel analysis. One stems from the econometric tradition, and focuses most explicitly on the problem of unobservables in the causal system (Baltagi, 2005; Frees, 2004; Hsiao, 2002; Woolridge, 2002). The analysis in this framework typically involves *pooling* or stacking the data across waves. This means that each row of data contains information on X and Y from a particular unit at only one of the panel waves, with information from unit (case) 1 at waves 1, 2, . . . through time *T*, followed in the dataset with information from case 2 at waves 1, 2, . . . *T* until the last row contains information on X and Y from the *N*th case at the last wave (T) of observation. The pooling procedure thus yields N*T total observations for analysis. This setup allows the researcher considerable power in estimating a variety of panel models, including those where Y_t is predicted by X_t (or X_{t-1}) as well as by an additional factor U that represents *unobserved* variables or influences on Y for a particular unit *i* that remain stable over time. These factors give rise to what is referred to in the econometric literature as *unobserved heterogeneity*. In these models, a single effect of X_t or X_{t-1} on Y_t is typically produced that is purged of the potentially confounding influence of the unobservables, with the information contained in the panel data being used in alternative models to "sweep out" or take into account the unobserved heterogeneity in different ways, and to model possible dependence in the units' idiosyncratic error terms over time.

Another approach to panel analysis stems from the structural equation modeling (SEM) and path analysis traditions in sociology and psychology (Bollen, 1989; Duncan, 1975; Finkel, 1995; Kenny, 1979; Kessler and Greenberg, 1981). In this framework, a separate equation for each dependent (endogenous) variable at each panel wave of observation is specified with a set of independent variables, which may themselves be either exogenous, or unpredicted by other variables in the model, or endogenous variables that are caused elsewhere in the overall causal chain. Thus Y observed at time 2 of the panel may be predicted by Y at time 1 (the "lagged endogenous variable") and a series of time 1 Xs, Y at time 3 may be predicted by Y and the Xs at time 2, and so forth. The resulting series of equations are then, given appropriate assumptions about error processes and the distribution of observed variables, typically estimated simultaneously through maximum likelihood or related methods in software packages such as LISREL, EQS, MLWin, or SAS. The SEM approach is often extended by including additional equations to model random measurement error in observed indicators of the exogenous and endogenous variables, resulting in a set of measurement equations linking *latent* variables with one or more error-filled indicators, and a set of structural equations linking the latent variables together in the presumed causal system. Such models may also be extended to test alternative causal lag structures in the model, such that variables may be presumed to exert causal influence on endogenous variables either simultaneously (i.e., at the same wave of observation), or lagged by one or more time periods. The SEM approach is particularly useful for the panel analyst in estimating a variety of reciprocal causal effects models that guard against the possibility of biases induced by measurement error in the variables, and that allow the flexible testing of alternative lag structures.

Until recently, panel researchers have typically conducted their analyses within one framework or the other, with the choice depending to some extent on the leanings of the particular social science or other discipline, and to some extent on the nature of the substantive problem and the threats to causal inference that were presumed to be especially serious in a given research area (i.e., measurement error versus unobserved heterogeneity). But much work has been done in recent decades to bring the two traditions together, or perhaps more accurately, to bolster the toolkit of each framework so that it incorporates some of the major strengths of the other. Thus the SEM approach has recently been extended to incorporate models of unobserved heterogeneity, while the econometric approach has expanded to include more extensive models of measurement error, reciprocal causality, and dynamic processes than were typically the case in decades past. While it would not be accurate to state that there are no remaining differences between the SEM and econometric approaches, it is nevertheless the case that panel researchers now have the tools to at least attempt to overcome the common threats to successful causal inference using either analytical framework.

In what follows, I shall provide an overview of the basic econometric and SEM methods used for linear panel analysis. I shall begin with the problem of unobserved heterogeneity, outlining the "fixed" and "random" effects models that deal with this problem. The discussion will then turn to the SEM approach for estimating dynamic models of reciprocal causality and then models with measurement error in the observed indicators. Finally, I will outline briefly more advanced models within the econometric and SEM traditions that attempt to incorporate both unobserved heterogeneity and dynamic processes in order to strengthen the causal inference process.[2]

2 Unobserved heterogeneity models for linear panel analysis

Consider a simple, single equation model of the effects of some independent variables X_1, $X_2 \ldots X_j$ on a dependent variable Y, each measured at **T** points in time on a sample of **N** individuals. There are no reciprocal effects specified and perfect measurement is assumed for all variables. We can then write the basic panel model for the relationship between the *J* number of X variables and Y for the *i*th case at a given *t* point in time as:

$$Y_{it} = \alpha + \beta_1{}^{*}X_{1it} + \beta_2{}^{*}X_{2it} + \cdots \beta_j{}^{*}X_{Jit} + \varepsilon_{it} \qquad (1)$$

This, of course, is the usual model analyzed in cross-sectional research (where T=1), using ordinary least squares (OLS) under the classical assumptions regarding the distribution of the idiosyncratic error terms ε_{it}. With panel data, though, it is clear that the classical assumptions regarding the error term are unlikely to be met, for the simple reason that the observations are not independent across time. That is, the error term for case *i* at time *t* is likely to be related to the error term for the same case at time $t+1$, $t+2$, etc. One strategy for handling this problem is to maintain the structure of equation (1) and estimate parameters with some variant of "robust" standard errors that

[2] This chapter will not discuss, except in passing, a third tradition for panel analysis, known variously as "hierarchical growth" or "mixed" models, or "latent growth" models. See Chapters 33–35 in this volume, Bollen and Curran (2005), and Singer and Willett (2002).

allow casewise dependence or other deviations from the standard assumptions.[3]

Most panel analysts, however, prefer to model the sources of the temporal dependence between observations of a given case more directly. One relatively simple setup is to assume that the dependence is produced by some stable, unobserved factor or factors **U** that are unique to a given unit and are also related to Y. This means that the "true" model (with U summarized as a single variable) is:

$$Y_{it} = \alpha + \beta_1{}^*X_{1it} + \beta_2{}^*X_{2it} + \cdots \beta_j{}^*X_{jit} + U_i + \varepsilon_{it} \quad (2)$$

For example, if Y is a country's extent of political repression and the Xs represent some measures of democracy and economic performance, the U term may encompass factors such as a country's "culture," history of violence, degree of ethnic homogeneity, and the like, all or some of which may not have been measured or have been otherwise available for inclusion in the analysis. If the model were at the individual level, such that Y represents, for example, a person's knowledge about politics, the U term may encompass intrinsic intelligence, motivation, or family socialization processes that produced individuals who are generally higher or lower on knowledge than would be expected from the values of the observed Xs in the model. The U_i are referred in the econometric literature as individual-specific, or unit effects.

In cross-sectional analyses, the unobserved stable U factor(s) are folded into the equation's unknown error term and the analyst can do little if anything about it. If the U are uncorrelated with the X_j observed variables, this would mean that the equation's explained variance is less than it might otherwise be, with inefficiency in the estimation of the standard errors for the β_j regression coefficients. But if (as seems likely), the U are related in some way to the observed $\mathbf{X}_j$, then the corresponding estimates of β_j will be biased. The potentially confounding effects of omitted variables is, of course, one of the most serious problems in nonexperimental research of any kind.

2.1 Fixed effects models

With panel data, however, at least some headway in attacking the problem can be made. Equation (2) may be rearranged to show that the presence of U implies that each unit has its own intercept $(\alpha + U_i)$, where U_i may be viewed as all the stable unit-level factors that lead that case to be larger or smaller than the overall average intercept (α) for the dependent variable in the sample. This suggests that one way of dealing with the problem with panel data would be to estimate equation (2) with OLS by including a dummy variable for N-1 units, with the coefficient on the dummies representing the individual-specific effects for each case (relative to an omitted baseline unit). This procedure is referred to as the LSDV (least squares dummy variables) method, and produces consistent estimates of the β_j coefficients for the X_j, controlling for stable unit effects that push the intercept for that case above or below the common (or baseline) intercept α. In this way we see that panel data can use the multiple observations on cases over time to begin to control for the effects of some kinds of variables that are not measured or observed in a given dataset.

The LSDV method is not usually applied, however, due to the need to include a potentially enormous number of dummy variables in large-N panel studies. A more common approach is to first express equation (2) in terms of the unit-level means of all observed variables, as in

[3] This is the motivating logic behind the class of panel estimators known as "generalized estimating equations (GEE)." See Hardin and Hilbe (2003) and Zorn (2001), and also Hardin and Hilbe, Chapter 31 in this volume.

$$\overline{Y}_i = \alpha + \beta_1 \overline{X}_{1i} + \beta_2 \overline{X}_{2i} + \cdots \beta_J \overline{X}_{Ji} + \overline{U}_i + \overline{\varepsilon}_i \quad (3)$$

This formulation is called the "between" equation, as all "within-unit" variation over time is averaged out, leaving a model that only considers variation between the individual units. Subtracting (3) from (2) yields the *fixed effects* (FE) model that sweeps the U terms out of the equation altogether:

$$\begin{aligned} Y_i - \overline{Y}_i &= \beta_1 (X_{1it} - \overline{X}_{1i}) + \beta_2 (X_{2it} - \overline{X}_{2i}) \\ &\quad + \cdots \beta_J (X_{Jit} - \overline{X}_{Ji}) + (\varepsilon_{it} - \overline{\varepsilon}_i) \qquad (4) \end{aligned}$$

Given that U_i is constant over time, $U_i = \overline{U}_i$ and hence drops out from equation (4). Estimating the resulting regression with OLS (with an appropriate adjustment for the standard errors due to the N number of unit means figuring in the estimation) yields consistent estimates of the β_j effects of the observed X_j variables.[4] The fixed effects estimator is also referred to as the "within" estimator, as it considers *only* variation in X and Y around their unit-level means, as all "between unit" variation (i.e., differences in $\overline{X}_i$ and $\overline{Y}_i$) are eliminated through the mean-differencing procedure.

A related approach to sweeping out the individual-level U_i is simply to lag equation (2) by one time period, as in:

$$\begin{aligned} Y_{it-1} &= \alpha + \beta_1{}^* X_{1it-1} + \beta_2{}^* X_{2it-1} + \cdots \beta_j{}^* X_{jit-1} \\ &\quad + U_{it-1} + \varepsilon_{it} \qquad (5) \end{aligned}$$

Subtracting this equation from (2) yields the *first difference* (FD) equation that also produces consistent estimates of the β_j:

$$\begin{aligned} Y_i - Y_{it-1} &= \beta_1 (X_{1it} - X_{1it-1}) + \beta_2 (X_{2it} - X_{2it-1}) \\ &\quad + \beta_J (X_{Jit} - X_{Jit-1}) + (\varepsilon_{it} - \varepsilon_{it-1}) \qquad (6) \end{aligned}$$

While both FD and FE methods are consistent, the fixed effect procedure is more commonly applied in multiwave panel data, as it makes use of *all* of the over-time variation in X in its calculations, and provides a more parsimonious method of expressing X and Y at a given time period as deviations from an overall mean value. Subtracting X_{t-1} in the first difference approach is inefficient in the sense that only one of the lag values of X is used to difference out the unit effect, and the one that is used is to some extent arbitrary (why not subtract X_{t-2} or X_{t-3}?).

The FE and FD procedures are relatively simple, yet powerful methods for panel models controlling for unobserved heterogeneity between cases. As Allison (1994) and Halaby (2004) have shown, these methods are also equivalent to the popular "difference in difference estimator" (DID) for assessing treatment effects in quasi-experimental and policy research. If, as is plausible in almost all nonexperimental research situations, there are unobserved differences between members of the treatment and control groups, then it may be the case that such differences and not the treatment itself may produce the observed differences between the two groups. Constructing the difference between both the treatment group's and control group's pre-test score and post-test scores, or, in multiwave data, between the treatment group's and control group's pre-test and post-test mean-deviated scores, essentially removes the unobserved (stable) differences between the two groups from consideration. The "difference in differences" between the treatment and control groups is then the pure effect of the treatment, controlling for these unobserved differences (and any

[4] That is, since N degrees of freedom are used in the calculation of the unit means, the appropriate df for the fixed effects model is NT-N-k-1. The constant term α in (2) may also be recovered by adding the grand mean for Y and each of the X_j to each of the deviation expressions in (4). Both of these adjustments are implemented automatically in STATA and other software packages for estimating econometric panel models.

observed covariates that are included in the model as additional controls).[5] Allison (1994) shows how this logic can be applied to estimating the effect of a variety of treatments with panel data, where the treatments may be given to individuals who were not randomly assigned to different conditions in quasi-experimental conditions, or where the "treatments" were simply the experience of some life event such as divorce, military service, unemployment, or job promotion between panel waves. Whenever stable, unobserved factors at the unit level may be related to the unit's likelihood of experiencing of these events and also related to the unit's value on some outcome variable, the FE or FD approach sweeps out the (stable) unobservables to produce consistent estimates of the effects of the event or treatment itself.[6]

Despite their clear strengths in eliminating the potentially confounding effects of unobserved hetereogeneity, the FE and FD models have certain features that render them problematic for some panel analyses. First, the differencing or mean-difference process eliminates not only the stable unobserved factors from consideration, but also all stable *observed* factors, such that the researcher can say nothing about the effects on the dependent variable of characteristics of units, countries, individuals, or other units that do not change over time. In many research situations, the effects of (nearly) stable individual-level factors such as education or sex, or, at the country level, stable political characteristics such as electoral systems or institutional arrangements are of prime theoretical interest, and FE/FD models are less attractive.[7] Second, as was noted above, the FE/FD method uses 1/T of the available degrees of freedom, which in short panels can be a relatively high cost. Third, it is also the case that the estimates of the unit-effect in short-term panel studies are based on only a few waves of observation, and hence may be unreliable to the extent that chance factors produce a few consistently high or low readings on the dependent variable for a given case over time. Since FE/FD methods take these unit-effects as given, they potentially overstate the "true" amount of temporal dependence produced by stable unobservables.[8]

[5] The logical equivalence of the DID estimator and the FE estimator is ensured only when a dummy variable for the wave of observation is also included in the FE model. This presents no special statistical problems and results in the "two-way" fixed effects model. This addition is necessary because the FE (and FD) models eliminate from consideration those observations that do not change on X over time; the inclusion of the time dummies thus captures the general time effect on Y that occurs for both treatment and control groups, with the additional impact of the treatment being given by its coefficient.

[6] It is also possible to test for the statistical significance of the unit effects as a whole through estimation of a nested F test that compares the R-squared of models with and without the inclusion of the N-1 unit effects.

[7] Interaction effects between time and stable unit-level factors may, however, enter FE models, so that the researcher may estimate how the impact of a stable variable changes across waves of the panel.

[8] It is also argued that FE methods are more appropriate in instances where the analyst wishes to make inferences conditioned on the observed units in the sample, as in analyses at the country level where, for example, Germany's unit effect is of interest and would be the same no matter how many different country samples are drawn. In instances where large numbers of individuals are sampled randomly from some population (and hence the specific individual effects are of less interest), the random effects approach to be discussed subsequently is arguably more appropriate. See Allison (1994) and Teachman et al. (2001) for counter-arguments on this point, claiming that FE and RE are simply alternative ways of dealing with the "nuisance" posed by the presence of the unit effects in (2) above.

2.2 Random effects models

An alternative approach to estimating equation panel models with unobserved heterogeneity, the *random effects* or *random intercepts* (RE) model overcomes some of these deficiencies, though not without some additional costs of its own. Consider the two sources of unobserved error in equation (2): the unobserved unit effect (Ui) and the unit-time-specific idiosyncratic error term ε_{it}. In the "random effects" approach, both of these sources of error are treated as the realization of random processes and assumed to be independent, normally-distributed variables (with variances denoted as $\sigma_u{}^2$ and $\sigma_\varepsilon{}^2$ respectively). Some units have a higher intercept on the dependent variable because of a large random U term, some units have a lower intercept because of a small random U term; added to this error is the idiosyncratic error ε_{it} which produces further deviations from the linear prediction of Y from the U and the B_jX_{jj} terms of the model. The only additional unknown in this setup compared to the pooled OLS panel model of equation (1) is $\sigma_u{}^2$, and thus one immediate benefit of the random effects model is that it saves a significant number of degrees of freedom compared to its FE or FD counterparts.

Estimation of the RE model proceeds under two assumptions. First, the two components of the composite error term, U_i and the ε_{it}, are assumed to be unrelated, otherwise no separate estimate of each would be possible. Second, and more important, both error terms must be assumed to be unrelated to the included X variables in the model, i.e., $E(X_{ji}U_i) = E(X_{ji}\varepsilon_{it}) = 0$. This is of course problematic, in that the setup assumes away the possible correlation between X and the unit-effect unobservables that prompts many panel analyses in the first place(!) The situation need not be so dire, however, as will be discussed in more detail below. Given these assumptions, the composite error term of the model, $U_i + \varepsilon_{it}$, has a fixed structure over time, with the variances (diagonal elements) being equal to $\sigma_u{}^2 + \sigma_\varepsilon{}^2$, and the covariances (off-diagonals) being equal to $\sigma_u{}^2$ for every time period. As such, estimation can proceed through feasible generalized least squares (FGLS) methods that weight the model by the inverse of the error variance–covariance matrix, in this case the weight θ calculated as:

$$\Theta = 1 - \frac{\sigma_\varepsilon^2}{\sqrt{T\sigma_u^2 + \sigma_\varepsilon^2}} \tag{7}$$

Once an estimate of θ is obtained through manipulation of the "within" and "between" regressions of equations (3) and (4) above[9], the model is then transformed as:

$$\begin{aligned} Y_i - \Theta\overline{Y}_i = \beta_1(X_{1it} - \Theta\overline{X}_{1i}) + \beta_2(X_{2it} - \Theta\overline{X}_{2i}) \\ + \cdots \beta_J(X_{Jit} - \Theta\overline{X}_{Ji}) + (U_{it} - \Theta\overline{U}_i) \\ + (\varepsilon_{it} - \Theta\overline{\varepsilon}_i) \end{aligned} \tag{8}$$

with this equation's error term (comprised of the two last terms in parentheses) now having constant variance and zero correlation across time units for each case.

Several features of the RE model are especially noteworthy. First, it can be seen that as θ approaches 1, it means that more and more of the composite error variance is made up of unit-level (between) variance $\sigma_u{}^2$. In the unlikely event that θ equals 1, all of the error variance is unit-level variance, and the RE estimator reduces to the FE mean-differenced estimator of the β. As θ gets closer to 0, more and more of the error variance is made up of the random $\sigma_\varepsilon{}^2$ component, with no unit-level variance to take into account, and the RE estimator thus reduces to the pooled OLS approach represented by equation (1). So the RE approach represents something of a middle ground between

[9] An estimate of $\sigma_\varepsilon{}^2$ is obtained from the "within" regression of (4), as all of the unit-level variation has been purged from this model. An estimate of $\sigma_u{}^2$ is obtained by manipulating the error term of the "between" regression of (3), which produces the error term $(\sigma_u{}^2 + \sigma_\varepsilon{}^2/T)$.

the FE and OLS models, weighted toward one or the other depending on how much of the error variance is comprised of unit-specific versus idiosyncratic components.[10] It may also be said that RE "adjusts" or "shrinks" the FE estimator back toward pooled OLS to the extent that the unit-level effects in general are either small, relative to overall error, or are unreliable due to a relatively small T (as can be seen from the denominator of equation (7)).[11]

Second, the RE model produces estimates of the effects of both changing *and* stable independent X variables, as its estimation equation (8) does not result in the elimination of any stable variable so long as θ is not 1 (or very close to 1, when estimates of stable variables will tend to be very imprecise). The ability to model Y as functions of both changing and unchanging X variables over time is one of the major advantages of the RE approach to unobserved heterogeneity; as noted, however, this (and other) advantages of RE may be enjoyed only to the extent that the assumptions of the model hold, i.e., that there is no correlation between the X_j and the U_i.

For this reason, there has been much debate in the panel literature over the applicability of FE versus RE. The "Hausman" test provides one way of adjudicating the dispute, by providing a test statistic to assess the significance of the difference between the FE and RE estimates. The logic behind the test is that, if the assumptions of the RE model hold, then FE and RE are two different ways to arrive at consistent estimates of the β, but RE is more efficient. If the assumptions of the RE model do not hold, then RE will be inconsistent, while FE will always be consistent. Thus, one should see similar estimates of the β_j if the RE assumptions hold, and different ones if they don't. Hausman (1978) showed that the statistic

$$\frac{\beta(FE) - \beta(RE)}{\text{var}\beta(FE) - \text{var}\beta(RE)}$$

is distributed as χ^2 with J degrees of freedom, with failure to reject the null hypothesis implying that the RE model is appropriate. If the Hausman null hypothesis is rejected, then it may be concluded that there is some violation of the RE assumptions and that a likely nonzero correlation between the X_j and the U_i exists which the analyst should take into account. One way to do so is through the FE or FD methods that eliminate the U_i from consideration in the estimation equation altogether. Another is by including in a random effects version of (2) the *unit-level mean* for each time-varying independent variable (i.e., $\overline{X}_j$) as additional predictors (Skrondal and Rabe-Hesketh, 2004, pp. 52–53). The RE "problem" may thus be viewed as an omitted variable issue, with the unit effects being potentially correlated (at the unit-level) with the included explanatory factors. Once the time-varying unit-level means are included, this correlation is essentially controlled for in the model, with the resultant composite error term satisfying the assumptions for FGLS estimation.[12]

2.3 Example: The effects of civic education on political knowledge in Kenya, 2002–2003

We illustrate these models with panel data collected on 401 individuals interviewed at three

[10] The proportion of total variance comprised of unit-specific variance is also referred to as the "intraclass correlation coefficient."

[11] It can also be seen from equation (7) that the RE and FE converge as T → ∞, so the issue of fixed versus random effects is not as relevant in large T panel (or time-series cross-sectional) studies.

[12] Plümper and Troeger (forthcoming) provide a similar method for incorporating unit-level means into a fixed effect model. In both cases, it is still necessary to assume that the time-invariant X_J are unrelated to $U_{i.}$. See Hausman and Taylor (1981) for an approach to this problem involving different assumptions about correlations between particular time-invariant X variables with the unit effects.

waves between February 2002 and June 2003 in Kenya, as part of a study evaluating the effects of attending civic education and democracy training workshops on changes in democratic knowledge, attitudes, and participation in the run-up to the Kenyan December 2002 presidential elections.[13] A total of 210 respondents were interviewed before they attended a civic education (CE) workshop between February and April 2002, with 191 individuals serving as the control group, as they were selected to match the "treatment" group on place of residence, gender, age, and educational status. There were two follow-up interviews for all respondents, one in November, some 7–9 months after the workshop, and another in April–May 2003, about one year after the initial workshop took place. Our concern here is whether the workshop exposure (the "treatment") led to significant changes in respondents' knowledge of politics, measured with four questions asking the name of various elected officials (the Vice President and the Provisional Commissioner) and various institutional provisions in the Kenyan political system (e.g., the length of the term of office of the President and the procedures for amending the Kenyan constitution).

Given that individuals selected themselves into the "treatment" group, in that attendance at local civic education workshops was purely voluntary, it is likely that the treatment group differs from the control group on some measured variables, such as interest in politics, and also on unmeasured factors that may influence political knowledge such as intrinsic intelligence, motivation, openness to political reform, personal discussion networks, and the like. To control for these potentially contaminating effects, a series of unobserved heterogeneity models were estimated with the three-wave data, and the results are shown in Table 29.1.

The model in the first column of the table shows the pooled OLS model of (1) above, i.e., one that contains no unobserved heterogeneity term whatsoever. In this model, individuals exposed to civic education workshops are on average .32 higher on the dependent variable after exposure than individuals in the control group, holding interest, education, sex, and age constant. The dummy variables for wave 2 and wave 3 show that, controlling for other independent variables, there is a .31 increase in knowledge in wave 2 compared to wave 1 (the baseline wave of the panel), and a .18 increase in knowledge in wave 3 compared to wave 1 for all individuals, including those in the control group. In the fixed effects model of column (2), the unit-specific effect specified in the theoretical model of (2) above is swept out through the mean-differencing process, leaving the "within" regression of individual deviations from their own means. In this model, the effect of civic education falls to .21, approximately two-thirds of its magnitude in the pooled model, though still significantly different from zero. Note that in the FE model, there are no estimates for the time-invariant factors of education, sex, and age, as they drop out of the model (along with U_i) through mean-differencing. Thus the FE model shows that exposure to civic education has a significant impact on later political knowledge, controlling for observed and unobserved stable factors at the individual level that may be correlated with both CE exposure and with knowledge.

An alternative random effects model is shown in column (3) of the table. As can be seen, the estimates are much closer to the pooled OLS model in column (1) than the fixed effects estimates, with CE now having a .31 effect on knowledge. Estimates of the stable observed control variables are also similar to their OLS values, as they should be, given the relatively

[13] For more information on the overall study, including details on the sampling and questionnaire design, see Finkel (2003), available at www.pitt.edu/~finkel.

Table 29.1 Unobserved heterogeneity models, Kenya three-wave civic education study

Variable	*(1)*	*(2)*	*(3)*	*(4)*
	Pooled OLS	*Fixed effects*	*Random effects*	*Random effects with unit-level means*
Civic education	.32	.21	.31	.21
	(.05)	(.10)	(.07)	(.10)
Political interest	.27	.13	.24	.12
	(.04)	(.05)	(.04)	(.05)
Education	.20		.20	.19
	(.02)	(dropped)	(.02)	(.06)
Male	.47		.48	.43
	(.05)	(dropped)	(.06)	(.06)
Age	−.005		−.005	−.005
	(.002)	(dropped)	(.002)	(.002)
Wave 2 dummy	.31	.41	.33	.41
	(.07)	(.08)	(.07)	(.08)
Wave 3 dummy	.18	.28	.20	.28
	(.07)	(.08)	(.07)	(.08)
Treatment mean				.13*
	----	----	----	(.13)
Political interest mean				.32
				(.09)
Constant	.78	2.01	.85	.33
	(.15)	(.13)	(.17)	(.21)
Adj. R-squared	.26	.10	.27	.28
Degrees of freedom	1193	796	1192	1190
Intraclass correlation	----	.46	.18	.18
Estimated θ	----	----	.22	.22

All coefficients statistically significant at $p<.05$ except *. Standard Errors in Parentheses. N = 401.

small estimated value for θ (.22). This model, however, assumes no correlation between the treatment or other observed variables and the random unit-level error term; a Hausman test shows that this assumption is likely to be violated, as the null hypothesis of no difference between the FE and RE models can be rejected ($\chi^2 = 13.88$, $p<.01$). Thus the choice here is between a FE specification, or a random effects specification with the unit-level means controlled (model 4). In both instances the CE effect is estimated to be .21; the difference between the models, aside from their assumption about the distribution of the U_i, is that the RE model also provides information on the effects of the time-invariant control variables that are differenced out of the FE specification.[14]

[14] It should be noted that the specification of the effect of CE in all of the models in Table 29.1 is that of an additive, permanent change, such that CE leads to an increase in knowledge in the immediate time period after exposure, with the effects persisting over time. Moreover, the effect of time itself is modeled with dummy variables for each time period. Allison (1994) shows other ways to model the impact of events, as well as alternative models of time effects. Random effects models that incorporate randomly-varying time-trends are also referred to as “hierarchical longitudinal growth models” (see Raudenbush and Bryk, 2002, and also Luke, Chapter 33 in this volume).

3 Dynamic panel analysis

To this point, it has been assumed that all of the temporal dependence in the model structure was rooted in the presence of the unit-specific effect (be it fixed or randomly distributed). That is, the only reason that a Y_i response at time t would be correlated with a Y_i response at time $t+1$ is the presence of the stable U_i term, which influences the response for the individual unit at all time periods. Nothing else in the models considered thus far suggests a dynamic process involving time-dependence in either the core of the model or in the idiosyncratic error terms. The longitudinal nature of the data, in other words, has been used primarily as a means to rid the model of the effects of the nuisance U_i term, and not to model any kind of dynamic processes *per se*. Controlling for unobserved stable unit effects is highly important for panel analysis, but it is often insufficient to take into account all of the temporal dependence in the data. Thus much of panel analysis is devoted to alternative means of modeling temporal dependence, either instead of, or in addition to, the heterogeneity models we have considered thus far.

3.1 Autocorrelated disturbances

One kind of additional temporal dependence is caused by correlations between the idiosyncratic error terms ε_i of successive panel waves. These *autocorrelated* disturbances could be the result of exogenous random shocks to the system that persist for several time periods, omitted variables that change over time, or correlated errors of measurement from one panel wave to the next. Such a model could be represented as in (2) above with the additional stipulation that the error terms are autocorrelated, as in:

$$Y_{it} = \alpha + \beta_1{}^*X_{1it} + \beta_2{}^*X_{2it} + \cdots \beta_j{}^*X_{jit} + U_i + \varepsilon_{it} \quad (9a)$$

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + \nu_{it}, \text{ where } \nu_{it} \sim N(0, \sigma_v^2) \qquad (9b)$$

This model is estimated in two stages. The first proceeds by estimating ρ through one of many available methods commonly used in time-series analyses (see Baltagi, 2005), and then weighting equation (9a) by the ρ estimate to produce a model with the well-behaved error term ν_{it}. In the second stage, either a fixed effects or random effects specification for the U_i is assumed, with estimation on the transformed equation (a) proceeding as in the models discussed earlier, either with mean-differencing (fixed effects) or estimation of the variance components of the error term and "θ-mean differencing" (random effects). One wave of observations is lost with the first-stage differencing procedure, so this model requires at least three waves of data.

3.2 Lagged endogenous variable models

An alternative model of temporal dependence in Y_i is perhaps more prevalent in the panel literature. In this model the Y_{it} response is determined by a series of X variables, either at time t and/or at a lag of $t-1$, along with the *lagged value of* Y_i, as in:

$$Y_{it} = \alpha + \beta_1{}^*Y_{it-1} + \beta_2{}^*X_{1it-1} + \beta_3{}^*X_{2it-1} + \cdots \beta_j{}^*X_{Jit-1} + \varepsilon_{it} \quad (10)$$

In this model, the lag value of Y (the "lagged endogenous variable") has a direct effect on the value of Y at the next time point, along with effects specified from prior values of X as well. We may include contemporaneous levels of X in the model as well, but for now we focus on the lagged effects for both the X_J and Y_{t-1}. The model captures the temporal dependence of adjacent responses on Y neither through their joint relationship with some stable unobserved U term, nor through the autocorrelation of adjacent unknown idiosyncratic errors ε_{it}, but rather because of the *direct influence* of Y at a given point in time on subsequent responses. With some simple algebra it can be shown that the model is equivalent to predicting the *change* in Y from its prior value, with the coefficient on lagged Y in the change-score version of the model being equal to ($\beta_1 - 1$):

$$Y_{it} - Y_{it-1} = \alpha + (\beta_1 - 1)^* Y_{it-1} + \beta_2{}^* X_{1it-1} + \beta_3{}^* X_{2it-1} + \cdots \beta_j{}^* X_{Jit-1} + \varepsilon_{it} \quad (11)$$

This means that, so long as the lagged dependent (endogenous) variable is included on the right-hand side, there is no difference between the expression of the model in terms of "static" scores in Y or as a "dynamic" change-score model in ΔY. Further, the effects of the X_j variables are exactly the same, regardless of the specification of (10) or (11).

The model of equation (10) or (11) differs from those we have considered thus far in several important ways. One is the absence of the unit-specific error term U_i, meaning that it is assumed that *all* of the temporal dependence of responses over time is due to the causal mechanism linking the lagged endogenous variable and the lagged (or contemporaneous) X_j to Y at any given point in time. This assumption can be relaxed, however, and we will examine models that include both lagged Y and the unit-specific effects below. But the most fundamental difference is the presence of lagged Y as a predictor in equations (10) and (11) in the first place, and the inclusion of this term that has generated a good deal of controversy in the panel literature.

All agree that whenever there are strong substantive reasons for assuming that prior values of a variable have a direct *causal* effect on its subsequent value, the dynamic model is entirely appropriate. For example, in models of attitude formation and change, it is often assumed that there is some natural "state-dependence," such that attitudes are determined directly by their prior values unless disturbed by some exogenous shock. Economic models of wealth may also assume that an individual's store of accumulated wealth *causes* subsequent levels through different investment decisions, employment, and educational opportunities, and the like. These processes would stand in contrast to variables that need to be "created anew" in each time interval, as, for example, behavior such as voting or political participation where engaging in the activity at one point in time may not necessarily "cause" subsequent behavior. In those cases, there may be little theoretical justification for estimating the dynamic model.

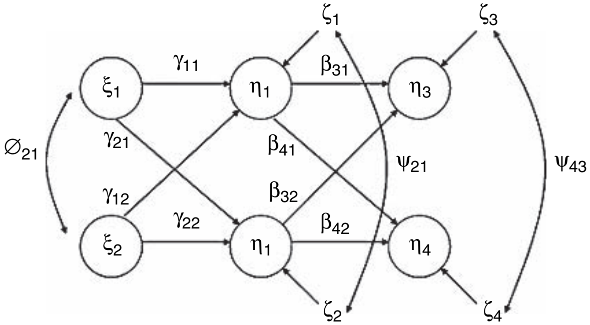
More controversial are the purely *statistical* reasons that have been advanced for including lagged Y in panel models. One is to serve as a proxy for unmeasured factors that lead to the response at both points in time, and many analysts argue that the heterogeneity models or autocorrelation models we have considered thus far deal more explicitly and more effectively with this problem (Allison, 1990; Liker, et al., 1985). Likewise, it has often been argued that including the lagged endogenous variable served as a statistical control for "regression to the mean" effects, whereby change in a variable is typically *negatively* related to its subsequent value (as seen in the ($\beta_1 - 1$) coefficient for Y_1 in equation 11). This occurs because high (low) initial values were likely to be the product of some random forces that are not likely to be repeated in subsequent observations. Others claim, however, that once random measurement errors are taken into account (with models that we consider below), regression to the mean is usually not sufficient to justify the inclusion of Y_{t-1} in the model. These arguments are far from settled in the literature, but they do point to the need for researchers to consider carefully the "epistemological status" of the lagged endogenous variable (Arminger, 1987), and not enter it automatically in panel models.

4 Structural equation panel models

The dynamic specification of (10) is a common starting point for panel analysis in the *structural equation modeling* (SEM) tradition. Instead of pooling the data across waves and estimating a single coefficient for each of the independent variables over the N*T units of observation, structural equation methods specify and estimate a system of equations, one for each dependent (endogenous) variable at each wave of observation. The overall model represents the interrelationships across all of the waves between the *exogenous* variables, i.e., those determined outside of the causal system and taken as “given,” and the *endogenous* variables, which are determined by other exogenous or endogenous variables in the model. In principle, each of the equations—i.e., the equation for Y in wave 2, Y in wave 3, etc.—could be estimated separately, but the SEM approach uses the information provided by the variances and covariances between *all* the observed variables in the model, even those that are not directly related to one another in any of the model’s equations, to allow the more efficient simultaneous estimation of all the model’s parameters. In panel models, this cross-wave covariation, along with the intrinsic temporal ordering between variables at earlier and later waves, provides the researcher great flexibility in estimating complex models with reciprocal causal linkages between variables, and, under some conditions, models that allow random measurement error in the observed indicators over time. These are the principal ways that the structural equation approach is used by panel analysts, with the dynamic specification of (10) most often at the core of these models.

The SEM approach may be illustrated with an example of the relationship between an individual’s partisan identification (i.e., the direction and strength of his or her attachment to the Republican or Democratic Party in the US), and presidential approval (i.e., whether he or she approves of the performance of the sitting President in office). In the political science literature, controversy rages over whether party identification is a stable characteristic which influences shorter-term perceptions such as candidate evaluations, presidential approval, or whether such short-term factors instead determine partisan change over time (e.g., Green and Palmquist, 1990). A structural equation model of these alternative processes is depicted in Figure 29.1, with the variables of partisan identification and presidential approval measured in three waves of observation in the National Election Panel Study of 2000–2002–2004. In order to facilitate the presentation through the chapter, the variables and coefficients are labeled with the LISREL nomenclature which is widely utilized within the SEM tradition (Jöreskog and Sörbom, 1994; Kaplan, 2000).

The model shows that party identification and presidential approval measured in 2000 are assumed to be exogenous variables, i.e., variables with causes outside the causal system, and are depicted as ξ_1 and ξ_2. Party identification and presidential approval in 2002 and 2004 are endogenous (η_1 through η_4), predicted by their own previous value and the previous value of the other variable.[15] That is,



ξ = Exogenous variable
η = Endogenous variable
β, γ = Regression coefficients
$\varnothing$ = Variance-covariances of the ξ exogenous variables
ψ = Variance-covariances of the ζ structural disturbances
ζ = Structural disturbance of the η

Figure 29.1 Three-wave, cross-lagged panel model

[15] All variables are coded in a “pro-Republican” direction so that high values on approval indicate either disapproval of a Democratic President (in 2000) or approval of a Republican President (in 2002 and 2004).

the model specifies a lagged effect from each variable on itself over time, and *cross-lagged* effects of party identification and presidential approval on each other. Each endogenous variable also has a random error term, depicted as ζ_1 through and ζ_4. The structural effect linking the exogenous variables ξ to endogenous variable η are labeled as γ coefficients, and the structural effects linking the endogenous variables to one another are labeled as β. Following common SEM practice, all variables are expressed as deviations from their mean, which eliminates consideration of the intercept in all of the structural equations.[16]

The four equations for the endogenous variables may therefore be written as:

$$\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1 \tag{12a}$$

$$\eta_2 = \gamma_{21}\xi_1 + \gamma_{22}\xi_2 + \zeta_2 \tag{12b}$$

$$\eta_3 = \beta_{31}\eta_1 + \beta_{32}\eta_2 + \zeta_3 \tag{12c}$$

$$\eta_4 = \beta_{41}\eta_1 + \beta_{42}\eta_2 + \zeta_4 \tag{12d}$$

or in matrix form as

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{13}$$

where
$\boldsymbol{\eta}$ = a vector of m endogenous variables (in this case 4)
$\mathbf{B}$ = an m by m matrix of β coefficients linking the m endogenous variables (here 4 × 4)
$\boldsymbol{\Gamma}$ = an m by n matrix of γ coefficients linking the n exogenous to the m endogenous (4 × 2)
$\boldsymbol{\zeta}$ = a vector of m structural disturbances for the endogenous variables (here 4)

Two other matrices are relevant from the point of view of specification and estimation: $\boldsymbol{\phi}$, the n by n matrix of the variances and covariances of the exogenous variables (here comprised of three distinct elements, the variances of ξ_1 and ξ_2 and their covariance, represented by the curved arrow labeled ϕ_{21}); and $\boldsymbol{\psi}$, the m by m matrix of the variances and covariance of the structural disturbances ζ (here 4 × 4, with the diagonal elements representing each equation's error variance and additional covariances estimated between the structural disturbances for n_1 and n_3 at wave 2, and for n_2 and n_4 at wave 3). These error covariances represent the residual covariance between party identification and presidential approval at a given panel wave that cannot be explained through the stability and cross-lagged effects in the model.

Several features of this model are important to note. First, it can be seen that each of equations (12a) through (12d) is simply a version of the dynamic panel model of (10) with the lagged values of Y and X as independent variables. Thus the estimate for γ_{11} in (12a) and β_{31} in (12c) represent the "stability" or "autoregressive" effect of party ID in wave 1 or 2 on its own value in wave 2 or 3; alternatively, subtracting 1 from these estimates will result in the effect of party ID at wave 1 or 2 on the subsequent *change* in party ID over the next panel wave. Similarly, the estimates γ_{12} and β_{32} represent the lagged effect of presidential approval in wave 1 or 2 on subsequent values of party identification, or, equivalently, on the change in party identification over time. The corresponding coefficients for presidential approval and party identification in equations (12b) and (12d) have exactly the same interpretation in terms of the respective variables on presidential approval or the

[16] LISREL and other SEM analysis packages do allow for the estimation of intercepts, and these kinds of models are widely used in the analysis of group differences (Sörbom, 1982) and in multilevel structural equation models (Muthén, 1994).

change in presidential approval over time. Thus the SEM setup here is designed to provide information on the relative stability of the two variables, as well as on the relative magnitude of the two cross-lagged effects across the waves 1–2 and waves 2–3 periods.[17]

Second, the model as specified contains no contemporaneous reciprocal linkages between party identification and presidential approval at wave 2 or wave 3, i.e., there are assumed to be no effects between η_1 and η_2 or between η_3 and η_4. This means that the cross-lagged model in Figure 29.1 is *recursive*, in contrast to a *nonrecursive* model that would contain feedback effects between variables observed at the same panel wave. The assumption of no contemporaneous effects may not be justified, as in many panel analyses the length of time for X to affect Y may be significantly shorter than the time lag between waves of observations. Nevertheless, in the absence of strong theory the cross-lagged model is usually a satisfactory initial model.[18] Nonrecursive models present more difficulties than recursive models in identifying and estimating causal parameters, as will be seen below.

Structural equation panel models are estimated in the same way as all SEMs (see Bollen, 1989; Kaplan, 2000). The variances and covariances between observed variables are expressed in terms of the unknown parameters γ, β, ψ, and ϕ, given the model's assumptions. In this case we assume that all variables are expressed in mean deviation form, that the ζ disturbances are unrelated to both the η and ξ that appear as independent variables in their respective equations, and that there are no covariances between any of the ζ disturbances.

Second, it is determined whether the model as a whole, and individual equations within it, are *identified*, i.e., whether there is sufficient information in terms of the observed variances and covariances to produce unique estimates of each of the model's parameters. All recursive models are either identified or overidentified (with more known variances and covariances compared than unknown parameters), though this will not be the case in the nonrecursive simultaneous effects models we will consider shortly. In the model of Figure 29.1, there are 17 unknowns—3 ϕ variances and covariances of the exogenous variables, 4 γ linking exogenous to endogenous variables, 4 β linking endogenous variables, and 6 ψ variances and covariances of the endogenous variables—and 21 known variances and covariances. This means that the model is *overidentified* with 4 degrees of freedom. When models are overidentified, there will be more than one solution for at least some of the model's unknowns, and this additional information can be used to assess how well the model fits the data as a whole.

Third, under the assumption of multivariate normality of the observed variables, maximum likelihood methods are typically used to estimate the model parameters. Intuitively, the ML procedures find the estimates of the unknown parameters, which, taken together, minimize the difference between the implied and actual variance–covariance matrices (see, e.g., Kaplan, 2000, pp. 24–27 for more details).

[17] It is important to note that the "stability" represented by the lagged dependent variable is not absolute stability in the sense of "no change," but rather stability in the sense of relative rankings of cases over time. When the autoregressive coefficients are closer to 1, this indicates that units with higher values at time (t) tend also to have higher values at time (t+1), though significant absolute changes may have occurred, either due to the effects of the other variables in the model or an overall change that affects all units equally.

[18] It is also the case that the cross-lagged model can be derived from the "continuous time" panel model where both X and Y exert continual influence on one another over time, as opposed to having effects in discrete time intervals corresponding to the waves of measurement (Coleman, 1968; Tuma and Hannan, 1994).

Finally, the variance–covariance matrix that is implied from the estimated coefficients is compared to the observed variance–covariance matrix to make assessments of the fit of the model as a whole. If a model can reproduce the observed variances and covariance very well, it is "consistent" with the data. If a model cannot, it is "inconsistent" with the data.

A variety of measures are available for assessing the significance of particular models in terms of their overall explanatory power and in terms of making comparisons between alternative models with different unknown parameters. For example, the quantity n^{*} log L_0, where L_0 is the likelihood function of the estimated model, is distributed as χ^2, with degrees of freedom equal to the number of estimated parameters. This "model χ^2", widely used as a measure of the overall fit of the model to the data with low values, relative to degrees of freedom, indicated better fit. Given the sensitivity of χ^2 to sample size, a variety of additional measures have been proposed to assess the explanatory power of a given model versus alternative baseline models; e.g., the Normed Fit Index (NFI) compares the model χ^2 to the χ^2 of a "complete independence" model with zero covariances among all variables in the population, while the Parsimony Normed Fit Index (PNFI) penalizes the NFI to the extent that the estimated model includes more and more parameters. See Hu and Bentler (1995) for a comprehensive overview of these issues. Finally, when models are "nested," such that one model can be defined by relaxing constraints on parameters in another model, the difference in χ^2 between the two models provides a test of the significance of the improvement in fit between the unconstrained versus constrained models.

The ML estimates of the cross-lagged model of Figure 29.1 for the NES panel data (N=738) are shown in Table 29.2. The model shows that there are significant cross-lagged effects in both directions between party identification and presidential approval, with the magnitude of the party-approval effect being approximately three to four times larger as the reverse. The model shows strong stabilities for the party and variable and weak stability for the approval of the President, especially between 2000 and 2002, which may be expected due to the change in Presidential administration (despite the coding changes described in footnote 12). Nevertheless, the model shows some support for the "revisionist" notion that short-term political evaluations influence the intensity of an individual's identification with particular political parties in the US. The estimates, however, are in the context of a poorly-fitting model, with a large and significant χ^2 and a relatively low Parsimony Normed Fit Index of .26.

Column (2) shows the flexibility of the SEM approach in terms of constraining particular parameters to be equal in order to test the relative explanatory power of alternative model specifications. In this model, the cross-lagged effects between party and approval from waves 1–2 and 2–3 are each constrained to be equal, thus gaining 2 degrees of freedom in the process. The results show nearly identical parameter estimates to model (1), and the difference in the two models' χ^2 is only .27, which, with 2 degrees of freedom difference, indicates that the unconstrained model does not fit the data significantly better than the constrained model. The PNFI shows a corresponding improvement to .39, reflecting the nearly equal explanatory power of this model with fewer estimated parameters. In practice, panel analysts will estimate a variety of alternative models, usually imposing equality constraints at the outset and relaxing them as necessary as indicated by χ^2 tests and other information about model fit. In this case, the overall indices show that neither of the models estimated provides a particularly good fit to the data, meaning that additional parameters, perhaps in the form of synchronous casual effects, may need to be included.

Table 29.2 Cross-lagged reciprocal effects models, party identification and presidential approval, American National Election Study, 2000–2002–2004

	(1)	*(2)*	*(3)*
	No equality constraints	*Equality constraints*	*Measurement error in party identification*
Stability effects			
Party identification, wave 1–2 γ_{11}	.80	.81	.97
	.79	*.80*	*.96*
Party identification, wave 2–3 β_{31}	.88	.88	1.04
	.83	*.82*	*.96*
Presidential approval, wave 1–2 γ_{22}	.15	.14	.09
	.16	*.15*	*.10*
Presidential approval, wave 2–3 β_{42}	.56	.57	.52
	.49	*.49*	*.44*
Cross-lagged effects			
Party to approval, wave 1–2 γ_{21}	.31	.31[a]	.37[a]
	.43	*.43*	*.49*
Party to approval, wave 2–3 β_{41}	.31	.31[a]	.37[a]
	.37	*.37*	*.43*
Approval to party, wave 1–2 γ_{12}	.17	.16[b]	.03[b*]
	.13	*.12*	*.02*
Approval to party, wave 2–3 β_{32}	.15	.16[b]	.03[b*]
	.10	*.11*	*.02*
Error covariances			
Wave 1 φ_{21}	2.25	2.25	2.26
	.62	*.62*	*.66*
Wave 2 ψ_{21}	.45	.45	.29
	.13	*.13*	*.09*
Wave 3 ψ_{43}	.44	.44	.25
	.10	*.10*	*.06*
Measurement error variance			.50
Party identification ε_1	----	----	
R-squared, party identification, wave 2	.77	.77	.95
R-squared, party identification, wave 3	.79	.79	.95
R-squared, presidential approval, wave 2	.29	.28	.31
R-squared, presidential approval, wave 3	.60	.60	.62
χ^2 (degrees of freedom)	133.33(4)	133.60(6)	10.91(5)
Normed fit index (NFI)	.97	.97	1.00
Parsimony normed fit index	.26	.39	.33

All variables statistically significant; standardized coefficients italicized.
Coefficients a, b constrained in models (2) and (3) to be equal. N = 738.

4.1 Alternative lag effects

Given enough waves of observation and enough knowns in the form of observed variances and covariances, the SEM approach allows considerable flexibility in specifying alternatives to the cross-lagged causal models of Figure 29.1. One possibility is a model with synchronous, or simultaneous impact of variables on one another at a given wave of observations. Such a model would be appropriate if the time lag for the causal influence of the independent variable is thought to be short, relative to the time period between observations. For example, in models of the impact of interpersonal networks, if the panel waves were separated by five years, a cross-lagged model may not capture network impact well, as its effect may either have dissipated in the intervening time period or the network itself may have changed since the initial panel wave, or both. In that case, a more accurate depiction of the causal process may be to include β effects between η_1 and η_2, and between η_3 and η_4 instead of the cross-lagged effects linking the variables from wave 1 to 2, and wave 2 to 3. It is also possible that both short- and longer-term causal lags could be present, in which case the additional β effects would be included along with the cross-lagged processes already specified.

In either case, the inclusion of synchronous causal effects yields a *nonrecursive* causal model whose estimation is significantly more complicated than in the recursive, cross-lagged only case. This is so for two main reasons. First, in synchronous effects models the assumption of no correlation between independent variables and the error term in their respective equations is untenable, meaning that the methods we have considered thus far would produce inconsistent estimates of the β. Imagine hypothetical causal arrows between η_1 and η_2 in Figure 29.1, with β_{21} representing the effect of party ID on approval in 2002, and β_{12} representing the reciprocal link between approval and party ID. It can immediately be seen that η_1 is related to ζ_2, since ζ_2 causes η_2 which causes η_1. Similar processes result in a nonzero correlation between η_2 and ζ_1, meaning that we cannot consistently estimate either β_{21} or β_{12}—or any of the other effects in their respective equations—with the information at hand.

Second, the inclusion of reciprocal synchronous effects raises the possibility that the model as a whole, or some of the individual equations, are not identified. For example, in a bivariate cross-section model with reciprocal causal effects, the model would not be identified because there would be four unknowns—the two β and the two structural disturbances—and only three known variances and covariances with which to estimate the unknown parameters.[19]

What is needed in the case of nonrecursive models in general, and synchronous effects panel models in particular, is more information in the form of additional observed variances and covariances. More information, of course, generates additional "knowns" so that the counting rule is more likely to be satisfied. But the additional variables must be related to the included variables in specific ways in order to be of use in solving the estimation problems posed in nonrecursive models. Specifically, what are needed are variables that can help satisfy the so-called *order condition* for identification, which states that if an equation involves m endogenous variables, then there must be at least $(m-1)$ *excluded exogenous variables* for the equation to be identified. That is, the equation must have at least one ξ that does *not* have an effect on the η endogenous dependent variable in question for every endogenous independent variable that does. For example,

[19] In that case, the model as a whole would not satisfy what is referred to as the *counting rule* for identification, whereby $(m+n)(m+n+1)/2 >= k$, with m and n representing the number of exogenous and endogenous variables, and k representing the number of free parameters.

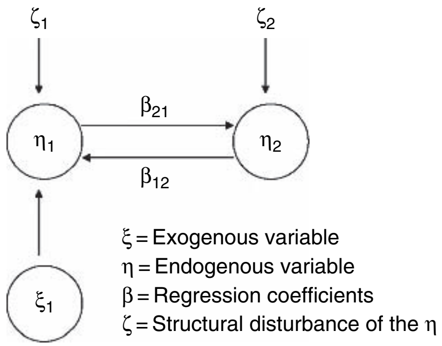


Figure 29.2 An instrument variable (ξ_1) in a nonrecursive causal model

in Figure 29.2 we add a variable ξ_1 to the equation predicting η_1 in a reciprocally-related bivariate cross-sectional model. Following the order condition, ξ_1 identifies the η_2 equation, as there is now one excluded exogenous variable from the equation and one included endogenous variable (η_1). Given that no exogenous variable is excluded from the η_1 equation, it is still not identified. Thus identification depends on the availability of additional exogenous variables that affect one and only one of the two reciprocally-related variables in a synchronous effects model.

The excluded exogenous variable, also called an *instrumental variable*, is then used to estimate the model parameters consistently. In the simplest case, such as depicted in Figure 29.2, the assumptions regarding $\xi_{1,}$ namely, that it is uncorrelated with the ζ and it has no direct causal effect on $\eta_{2,}$ allow the expression of β_{21} as:

$$\text{Covariance}(\xi_1, \eta_2)/\text{Covariance}(\xi_1, \eta_1)$$

with the covariance between η_1 and the structural disturbance ζ_2 no longer contaminating the estimate of the β_{21} causal effect. In more complex cases, there may be more than one instrument available for inclusion in the model, in which case estimation proceeds through "two-stage least squares" (TSLS) methods or through the general maximum likelihood procedures discussed above. In the widely-utilized TSLS setup, the endogenous independent variable is first regressed on *all* exogenous variables, including all of the instruments, generating a "predicted η" which is uncorrelated with all of the model ζ. Then, in the second stage, the dependent variable is regressed on the exogenous independent variables and the "predicted η" from the first stage (with appropriate corrections to the standard errors in the second stage).

All of these procedures, however, depend on the satisfaction of the assumptions of instrumental variable analyses. That is, there must be variables included in the model that affect *one and only one* of the two reciprocally-related endogenous variables, and these variables must be exogenous in the sense of unrelated to the structural disturbances of the endogenous variables. These variables are difficult to find in many practical research situations. In panel designs, however, it may be reasonable under some conditions to assume that the lagged values of variables are exogenous and related to the endogenous variables in such ways as to facilitate identification and estimation. For example, consider a "pure" two-wave synchronous effects version of Figure 29.1 with no cross-lagged effect from either ξ_1 to η_2 or from ξ_2 to η_1. In that case ξ_1 would be used to identify the η_2 equation and, similarly, ξ_2 would be used to identify the η_1 equation. Thus, if the assumptions of exogeneity can be justified, the longitudinal structure of panel data can provide additional information in the form of instrumental variables for purposes of identifying and estimating nonrecursive, reciprocal effects models.

In many instances, however, the situation will be complicated by the violation of the exogeneity assumptions. If there are autocorrelated disturbances present in the model, for example, then the lagged value of a variable will *not* be unrelated to the structural disturbance of the endogenous variables and hence will not be a suitable instrument. Further, in models with *both* cross-lagged and synchronous effects, the lagged value of a variable, even if unrelated to the structural disturbances, is assumed to have effects on both endogenous variables at subsequent waves of observation, and so cannot be used to identify the subsequent wave's equations. One possible solution would be to include additional exogenous variables as instruments, provided that they satisfy the exclusion restrictions discussed above. Another common possibility is to identify the model through the use of equality constraints, such that the cross-lagged effects may be assumed to be equal from waves 1 to 2 and 2 to 3, the synchronous effects may be assumed to be equal in waves 2 and 3, and perhaps the stability effects equal across waves as well. Given that some of these models will be nested within each other, comparison of chi-square goodness of fit measures can provide some insight into the models that are most consistent with the observed data; with that information, along with the parameter estimates found for the various models, the analyst may arrive at conclusions regarding the likely pattern of lag causal effects between variables over the course of the panel observations.

In the current example, we re-estimate model (2) in Table 29.2 by including both synchronous and cross-lagged effects between party and approval, and specifying equality constraints between the two sets of contemporaneous effects. The results show that the cross-lagged effects are still significant in both directions, while neither contemporaneous effect is significant, with estimated values of .001 from party to approval and −.05 from approval to party. The model χ^2 is 133.3 with 4 degrees of freedom, thus not a significant improvement from the cross-lagged only model in Table 29.2, given the loss of 2 degrees of freedom. We conclude that the cross-lagged model is superior to a model with both cross-lagged and synchronous effects, though neither model fits the data well.

4.2 Measurement error

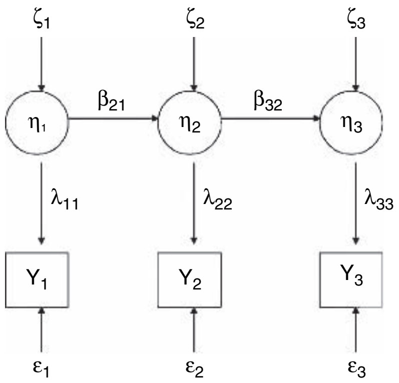
The SEM approach to panel analysis is also often extended to include the estimation of causal effects, controlling for errors of measurement in the variables of interest. It is well-known that random measurement in independent variables causes estimation of regression coefficients to be biased, downward in the case of bivariate equations and in unknown directions in multivariate models (Bollen, 1989). In cross-sectional analyses, there is often not enough information to identify and estimate both the structural effects *and* the measurement error that may be present in the model, as there needs to be multiple observed indicators of the presumed "latent" (error-free) variable of interest in order to proceed. With panel data, though, the information that is provided from the same variables over time allows much more flexibility in estimating structural effects once measurement error is taken into account, and in estimating and assessing the measurement properties of particular indicators as well. In the LISREL and other applications of the SEM framework, all of these effects are estimated simultaneously, providing powerful additional tools to the panel analyst in strengthening the causal inference process.

Figure 29.3 shows a three-wave autoregressive panel model that includes a random measurement component. The model depicts each wave's indicator of y as a function of an unobserved latent "true" score η plus a random measurement component ε. In equation form, the model is written in two parts as:

$$y_{it} = \lambda_{kt}\ \eta_{jt} + \varepsilon_{it} \quad \text{“Measurement Model”} \quad (14a)$$

$$\eta_{it} = \beta_{t,t-1}\ \eta_{jt-1} + \zeta_{it} \quad \text{“Structural Model”} \quad (14b)$$

with the ε assumed to be normally distributed random variables, uncorrelated with the η and the structural disturbances ζ, and, in basic models, uncorrelated with each other over time. The



η = Endogenous variable
ζ = Structural disturbance of the η
β = Regression coefficients
y = Observed indicator
ε = Measurement error
λ = Measurement model factor loading

Figure 29.3 Three-wave autoregressive model with measurement error

model differs from (13) because of the presence of random measurement error in the observed variables, which was assumed (unrealistically) to be zero in all of the models considered to this point. Expressing, for example, the structural portion of the model from wave 1 to wave 2 in terms of the fallible indicators shows the consequences of this omission, setting $\lambda = 1$ for simplicity:

$$y_{2i} = \beta_{21}\, y_{1i} - \beta_{21}\varepsilon_{1i} + \zeta_{2i} + \varepsilon_{2i} \tag{15}$$

As can be seen, the regression of y_2 on y_1, with both variables containing random error, will have a larger error variance than the true structural disturbance ζ_2, with a corresponding lower R-squared and inefficient estimates of the model coefficient's standard errors; even more consequential is that estimates of β_{21}, the autoregressive or "stability" effect in the model, will be inconsistent, as y_1 is intrinsically related to the error term in equation (15) due to the presence of ε_1.

This is an important result, showing that measurement error in the basic dynamic panel model yields an incorrect estimate of the stability parameter β, often lower than its true value. And to the extent that other independent variables are (positively) related to both y_1 and y_2, their effect in equation (15) is likely to be biased *upwards*, leading to potentially erroneous conclusions about their effects on y_2 or Δy. Thus, it is essential to take measurement error into account in panel models, and failure to do so is likely to negate many of the advantages of panel analysis in estimating dynamic causal processes.

Given the statistical nature of the problem, i.e., the correlation between the independent variable y_1 and the error term in equation (15), one solution would be to use instrumental variable analysis. If an exogenous variable ξ exists, such that it was uncorrelated with both the measurement error in y and the structural disturbance ζ, then estimation could proceed via the TSLS or the ML procedures considered thus far. In the first stage, y_1 would be regressed on all exogenous variables, including the instrumental variable(s), and in the second stage, y_2 would be regressed on all independent variables in its equation, along with the predicted value of y_1 from the first stage, which would be purged of the correlation with its error term. There are two drawbacks to this approach. One is practical, in that it is commonly difficult to find exogenous variables that satisfy the exclusion restriction discussed above, i.e., that are related to a fallible indicator in one wave of observation but not the next.

Second, the IV solution, even if successfully implemented, does not provide information about the measurement properties of indicators in the model, which may often be of considerable interest. In terms of the measurement model of (14a), we may wish to know the "reliability" of y, defined as the quantity

$$\text{Variance}(\eta)/\text{Variance}(y)$$

or the proportion of the observed variance that is comprised of "true score" variance. This information may be especially useful in multiple indicator models, i.e., those where more than one indicator for a given latent construct is available. In that case, assessing the reliabilities of the specific indicators, both absolutely and in relation to one another, provides additional information that can be used to determine the adequacy of the indicators or their suitability as measures of the underlying construct.

As the amount of information available increases, in terms of waves of observation and number of indicators of latent variables, panel models with measurement error may be identified and estimated with fewer and fewer restrictive assumptions. In the model of equations (14a) and (14b), for example, we have six observed variances and covariances, and 11 unknowns—the three λ, the three variances of the measurement error ε, two β, and three variances of the structural disturbances ζ. We may make some progress in setting all of the λ to equal 1; this has no substantive bearing on the model and simply puts the latent variable on the same measurement scale as the observed y indicators.[20] This leaves 8 unknowns and 6 knowns in the model. Wiley and Wiley (1970) propose that identification be achieved in this case by constraining the variance of the measurement error term ε to be equal across the three waves, gaining 2 degrees of freedom in the process as well. Under these assumptions, the model is just-identified and the relevant parameters can be solved through algebraic manipulation of the observed indicators' variances and covariances, as:

$$\begin{aligned}\text{var}(\varepsilon) = \text{var}(y_2) - [&\text{Covariance}\\ &(y_3, y_2)^{*}\text{Covariance}(y_1, y_2)/\\ &\text{Covariance}(y_1, y_3)] \qquad (16a)\end{aligned}$$

$$\begin{aligned}\beta_{21} = \text{Covariance}(y_1, y_2)/[&\text{Variance}(y_1)\\ &- \text{Variance}(\varepsilon)] \qquad (16b)\end{aligned}$$

$$\beta_{32} = \text{Covariance}(y_1, y_3)/\text{Covariance}(y_1, y_2) \qquad (16c)$$

Once these manipulations are accomplished, it is straightforward to solve the reliabilities of the indicators, given the observed variance $(y_t) = \text{var}(\eta_t) + \text{var}(\varepsilon)$.[21]

More generally, the maximum likelihood estimation of measurement error and structural coefficients may proceed simultaneously within the LISREL or other applications of the SEM approach. The measurement equations in (13a) would be summarized in matrix form as:

$$\mathbf{y} = \mathbf{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \qquad (17a)$$

and, if measurement error were assumed to be present in the exogenous ξ variables, then

$$\mathbf{x} = \mathbf{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \qquad (17b)$$

with the $\mathbf{\Lambda}$ matrices being (q × m) and (p × n) matrices of the λ linking the q indicators (y) of the m endogenous variables and p indicators (x) of the n exogenous variables, and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ being (q × 1) and (p × 1) vectors of the measurement errors in y and x, respectively. The implied variance–covariance matrix that expresses the knowns in terms of the unknown model parameters can then be expanded to include the

[20] In multiple indicator models, one indicator's λ is set to 1 for the same reason, while the other λ are free to vary.

[21] The single-indicator, three-wave model may also be identified through the Heise (1969) procedure. In this model, the latent and observed variables are standardized, so that the unknowns are the two β stabilities and the three λ coefficients, which represent path coefficients linking the latent variables and observed indicators (and thus λ^2 represents the reliability coefficient). Under the assumption of equal reliabilities across waves, the model is just-identified.

unknowns in $\Lambda_y \Lambda_x$, and the variances of the ε and δ, denoted as $\boldsymbol{\theta}_y$ and $\boldsymbol{\theta}_x$, respectively (see Kaplan, 2000, p. 56). Maximum likelihood estimation then proceeds as before, assuming multivariate normality for all the observed variables (and assuming that appropriate constraints are imposed when necessary to achieve parameter identification).

There are several options for incorporating the information regarding measurement error into full-blown cross-lagged or synchronous structural equation models. In three-wave single-indicator models, one method is to fix the measurement error variances of the indicators at the values that are obtained through the Wiley-Wiley procedure, with the structural effects then estimated while correcting for the unreliability of the indicators. Another method is to allow the measurement error variances to be completely free parameters, with LISREL or alternative programs producing simultaneous estimates of measurement and structural coefficients in the model.

This procedure is illustrated in model (3) of Table 29.2, which shows a reanalysis of the cross-lagged panel model of Figure 29.1 while allowing for measurement error in the party identification equation. As can be seen, the stability of the party variable rises considerably compared to the previous estimation, and neither cross-lagged effect from approval to party identification now is statistically significant or substantively meaningful, with standardized values of .04 or less. By contrast, the cross-lagged effect from party to approval is somewhat larger than in the no-measurement-error model, and the overall model χ^2 is much improved (10.91 with 5 degrees of freedom, compared with 133.6 with 6 degrees of freedom in the previous model). Thus interpretation of the causal effects in panel models may be substantially altered once measurement error is taken in to account; in this case the measurement error model shows results that are much more in accord with traditional views of party identification as the "unmoved mover" of short-term political evaluations (Green and Palmquist, 1990). And in this case, the reliabilities of the party identification measures are calculated as approximately 90%, meaning that only 10% of the indicator variance was estimated to be "error;" when indicators exhibit less reliability, then differences between measurement error models and models assuming perfect measurement will be even greater.[22]

It should be emphasized, however, that the measurement models estimated thus far depend on their own set of assumptions that need to be justified. For the Wiley-Wiley procedure, for example, it must be assumed that the error variances are equal over time for the three-wave panel model to be identified. This may be unrealistic, and work conducted with longer-term panels suggests that error variances tend in fact to shrink over time (Jagodzinski, et al., 1987). In the present case, moreover, the Wiley-Wiley method failed to produce credible estimates in the case of presidential approval, indicating that the assumptions in the model were unlikely to be satisfied.

In such instances, the analyst may simulate results by plugging in different values of error variances, assuming perhaps some degree of constant shrinkage over time. But the more promising solution, as always in measurement error models, is whenever possible to add waves of observation and/or additional indicators for the latent constructs. With four waves of data, the assumption of equal error variances can be relaxed such that the measurement errors of the "inner" indicators y2 and y3 are identified without constraint; moreover, the stability effect from wave 2 to wave 3 (β_{32}) is overidentified, thus indicating that the fit of the model as

[22] For example, the observed variance of y_2, party identification in 2002, is 4.84. Given the estimated error variance of .5, this yields a reliability estimate of $(4.84 - .50)/4.84 = .90$.

a whole can be assessed. With longer-wave panels, more parameters will be identified without the restrictive constraints of the Wiley-Wiley procedure. And when multiple indicators are available, models that incorporate autocorrelation between the errors of measurement over time may also be estimated, with fewer constraints necessary as the waves of observation and number of indicators increase.

5 Dynamic panel models with unobserved heterogeneity

A natural extension of the models that we have been considering thus far is to incorporate both dynamic causal processes (and potentially measurement error) along with unobserved heterogeneity into a single model. Both the econometric and the SEM panel traditions provide the ability to incorporate and estimate these kinds of models.

The dynamic model with heterogeneity takes the following general form:

$$Y_{it} = \alpha + \beta_1{}^* Y_{it-1} + \beta_2{}^* X_{1it} + \beta_3{}^* X_{2it} + \cdots \beta_j{}^* X_{jit} + U_i + \varepsilon_{it} \qquad (18)$$

with β_1 representing the effect of the lagged endogenous variable Y_{t-1} and U_i, as before, representing the unit-specific effect. It can be seen that the model combines equation (2), the basic model for unobserved heterogeneity, with equation (10), the basic dynamic model with contemporaneous effects of the Xs, though lagged values of X could also have been included in the model.[23] The combined model thus corresponds to a situation where, for reasons discussed earlier, the lagged endogenous variable is thought to exert direct causal influence on its subsequent values, *and* there are stable unobserved differences between the units that push individual cases higher or lower on the dependent variable, aside from the included Xs and aside from the dynamic processes represented by the effect of Y_{t-1}.

The inclusion of both kinds of effects in the same panel model results in some estimation difficulties, however. The problem stems from the fact that Y_{t-1}, the lagged endogenous variable, is intrinsically related to the composite error term of (18) due the presence of U_i. This can be seen by lagging equation (18) by one time period, as in:

$$Y_{it-1} = \alpha + \beta_1{}^* Y_{it-2} + \beta_2{}^* X_{1it-1} + \beta_3{}^* X_{2it-1} + \cdots \beta_j{}^* X_{jit-1} + U_i + \varepsilon_{it} \qquad (19)$$

which shows the direct dependence of Y_{it-1} on U_i, and hence the bias produced by traditional methods of estimating (18). Moreover, one popular method for “sweeping out” the unit effect, the “fixed effect” transformation of equation (4), fails to correct the problem, as the solution for eliminating U_i from consideration produces a transformed error term of $(\varepsilon_{it} - \bar{\varepsilon}_i)$. Since $\bar{\varepsilon}_i$ contains some portion of ε_{it-1}, then the lagged endogenous variable (Y_{t-1}) is still related to the mean-differenced error term, and thus in relatively short panels the biases in estimating the dynamic effects in this model still exist.[24]

The solution to the problem lies in an adaptation of the first-difference (FD) model considered above in (6):

$$\begin{aligned} Y_{it} - Y_{it-1} = \alpha + \beta_1{}^*(Y_{it-1} - Y_{it-2}) \\ + \beta_2{}^*(X_{1it} - X_{1it-1}) + \beta_3{}^*(X_{2it} - X_{2it-1}) \\ + \cdots \beta_j{}^*(X_{jit} - X_{jit-1}) + (\varepsilon_{it} - \varepsilon_{it-1}) \end{aligned} \qquad (20)$$

[23] As will be seen, in the SEM version of the model, the lagged values are normally the values that are included in order to ensure model identification.

[24] As $T \rightarrow \infty, \bar{\varepsilon}_i \rightarrow 0$, and hence the error term in (18) would no longer be related to Y_{it-1} so long as no autocorrelation is present in ε_{it}. But with short panels, bias on the order of 1/T exists. See Nickell (1981).

In this model, the U_i have been differenced out of the equation, with the resulting error term the difference between the idiosyncratic error at times t and $t-1$. It can be seen, however, that the *differenced* lagged endogenous variable $(Y_{it-1} - Y_{it-2})$ will still be related to the differenced error term, given the presence of ε_{it-1} in the latter.

What are needed are instrumental variables that are related to the differenced lagged endogenous variable but unrelated to the differenced error term, and various candidates have been proposed in the literature. Anderson and Hsiao (1982) suggest several possible instruments, one being the *twice-differenced* lagged endogenous variable $(Y_{it-2} - Y_{it-3})$, and the other being the *level* of Y_{it-2}; both proposed instruments are unrelated to the error term in (20).[25] One drawback to the former solution, however, is that it requires at least four waves of data, and subsequent work also suggests that the twice-differenced lagged endogenous variable is often a poorly-performing instrument in that it is usually only weakly related to $(Y_{it-1} - Y_{it-2})$.

The Arellano-Bond solution rests on the fact that the panel structure of the data provides more and more potential instruments in equation (20) as the number of waves of observation increases (Arellano and Bond, 1991). For three-wave data, Y_{it-2} may be used as an instrument for $(Y_{it-1} - Y_{it-2})$, for four waves of data Y_{it-2}, Y_{it-3} and $(Y_{it-2} - Y_{it-3})$ may be used, for five waves of data Y_{it-2}, Y_{it-3}, Y_{it-4} and all the respective changes scores may be used, and so on. So as one moves through the panel more and more instruments are included to arrive at more precise estimates of the dynamic and other effects in the model. In the Arellano-Bond formulation, the various lagged levels and differences of the exogenous X variables are also included as instruments for the same reason. One drawback to the method is its inapplicability when there is serial correlation in the original equation's (18) idiosyncratic error term, and the need for at least four waves of data to test this assumption.[26]

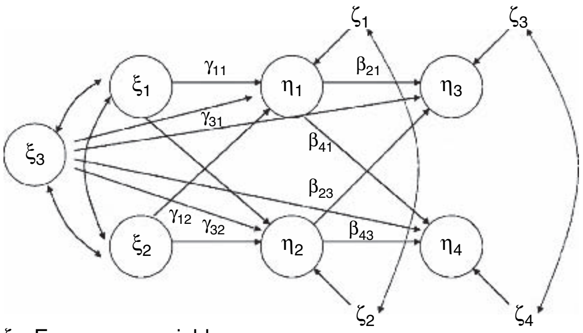
The strategy of using the panel structure of the data to find suitable instruments is also employed in econometric models that allow for reciprocal causality and measurement error. As has been discussed above, the statistical problem that results from either reciprocal causal effects specification, or from the presence of measurement error in variables, is an intrinsic correlation between the independent variable(s) and the structural disturbance term for that variable's equation. Hence some of the independent variables X in either (2), the original unobserved heterogeneity model, or the dynamic model (18) must be treated as endogenous. This leads naturally to the application of instrumental variables analyses in the context of first difference or fixed effects models, using lagged values of the independent variables as instruments under certain conditions (see Halaby, 2004, pp. 532–535; Woolridge, 2002, Chapter 12).

Dynamic models with unobserved heterogeneity may also be estimated within the SEM framework, though applications of this kind are less common in the literature (Dorman, 2001). An example of such a model is shown in Figure 29.4, with the basic features being the dynamic cross-lagged reciprocal effects specification considered earlier, along with an additional exogenous variable (ξ_3) that represents the individual-specific effect for each unit.

The unit-effect corresponds to the U_i term in equation (18), or, if no dynamic processes are specified, to the U_i term in the basic unobserved heterogeneity model of equation (2). It

[25] This strategy would not have worked with the fixed effect transformation, as Y_{it-2} would still be related to the portion of the error term $\bar{\varepsilon}_i$ which contains ε_{it-2}.

[26] See Wawro (2002) for discussion of alternative dynamic panel estimators, and application of these procedures to the controversy regarding the endogeneity of political party identification.



ξ = Exogenous variable
η = Endogenous variable
β, γ = Regression coefficients
ζ = Structural disturbance of the η

Figure 29.4 Three-wave cross-lagged model with stable unobserved variable (ξ_3)

is assumed to be stable over time (and thus can be represented by a single ξ), and it is assumed to be related to each of the other latent variables in the model as well. The unit-effect is specified within the SEM framework as an additional latent variable *with no observed indicator* and with variance set arbitrarily to 1—i.e., it is a "phantom variable" that is identified only if there is enough other excess information from the observed variances and covariances in the model. In the present case, the model is identified so long as some equality constraints are placed on the coefficients in the equations for the wave 2 and wave 3 variables, e.g., equal cross-lagged effects, equal stabilities, or equal error variances. Of course, alternative models that impose all of these constraints may also be estimated and compared. In the present example, an unobserved variable model linking party identification and presidential approval over time shows an excellent fit to the data (chi-square of 3.66 with 3 df), and the results indicate that there are no significant cross-lagged relationships between the two variables in *either* direction.

There are several attractive features of the SEM version of this model. First, the analyst may make use of the full range of SEM-related procedures for incorporating measurement error into the analysis, as the identification of the structural and measurement portions of the model are largely independent. Indeed, comprehensive models of this kind may often be necessary to estimate, as it otherwise may be unclear whether the unobserved variable represents the stable "unit-effect" or simply a latent variable that represents one or more of the observed constructs purged of measurement error. Second, alternative unobserved variable models may be tested and compared in terms of their ability to account for the observed data. This is especially useful in that a model with the unit-effect being related to only the endogenous variables over time will be nested within a model that has the unit-effect related to both endogenous and exogenous variables.[27] In this way the covariance structure analysis can provide a statistical test of the random effects versus fixed effects specification of the unobserved heterogeneity model (see Teachman, et al., 2001, for further development of this model). Finally, with enough waves of observation, more elaborate unobserved variables models may be specified, some that allow the unobserved variable to change over time (Dorman, 2001). In these ways the full power and flexibility of the SEM approach can be used to estimate models providing comprehensive attempts to overcome the most significant threats to successful causal inference in nonexperimental research.

6 Conclusion

This chapter has outlined two approaches to panel analysis, one focusing on the problem of unobserved heterogeneity, and the other focusing on dynamic causal processes

[27] That is, in one model the covariances between the unit effect and any other ξ would be fixed at zero, while in the other they would be estimated parameters.

and measurement error. Models incorporating unobserved heterogeneity were described mainly in the context of pooled econometric-type estimation, while the dynamic models with measurement error corrections were described mainly in the context of structural equation modeling procedures. As mentioned, however, recent work within the two traditions has resulted in a greater convergence of models and analytic strategies. This convergence is likely to continue, as newer developments in the field are even more explicitly synthetic in their approach.

For example, the recent work of Skrondal and Rabe-Hesketh (2004) incorporates all manner of latent variables, from unobserved heterogeneity to "true score variables" purged of measurement error in their indicators, to latent responses that represent missing values of partially-observed variables, into a single analytical framework. More generally, the realization that panel and "multilevel" data share the same logical structure (as the observations over time are nested within individual units) has led to the development of models that incorporate intratemporal growth processes at the "lower" level that may also vary randomly at the "higher" level. Such models incorporating randomly-varying intercepts and randomly-varying slopes may be estimated either through an extension of the random effects model discussed above, or with structural equation methods that treat the intercepts and slopes as "latent" variables that vary randomly across units (Bollen and Curran, 2005; Singer and Willett, 2002; Chapters 33–35 of this volume). Future developments in linear panel analysis, then, promise to build on the approaches presented in this chapter, to synthesize and to extend them in important new directions.

Acknowledgement

I thank George Krause for advice and encouragement, and Andrea Castagnola for exceptional research assistance.

Data and software

STATA 9.0 was used to estimate the pooled "econometric"-type models in the first section of this chapter. This software is available at: http://www.stata.com/. Many other software packages can be used for these models as well, including Mplus (http://www.statmodel.com/), SAS(http://www.sas.com/), MLwiN (http://www.cmm.bristol.ac.uk/), SPlus (http://www.insightful.com/products/splus/default.asp), and LIMDEP (http://www.limdep.com).

LISREL 8.72 was used to estimate the structural equation models in the second section of the chapter. This software is available at: http://www.ssicentral.com/lisrel/index.html. Other popular packages available for this kind of analysis are Mplus (http://www.statmodel.com/), SAS (PROC CALIS) (http://www.sas.com/), SPSS (AMOS) (http://www.spss.com/amos/), EQS (http://www.mvsoft.com/), MX (http://www.vcu.edu/mx/) and Smart Plus (http://www.smartpls.de/forum/).

The data used to estimate all models in this chapter can be found at www.pitt.edu/~finkel/data.htm.

References

Allison, P. (1990). Change-scores as dependent variables in regression analysis. *Sociological Methods and Research*, 20: 93–114.

Allison, P. (1994). Using panel data to estimate the effect of events. *Sociological Methods and Research*, 23: 179–199.

Anderson, T. W. and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18: 47–82.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58: 277–297.

Arminger, G. (1987). Misspecification, asymptotic stability, and ordinal variables in the analysis of panel data. *Sociological Methods and Research*, 15: 336–348.

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data*. London: Wiley.

Beck, N. and Katz, J. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, 89: 634–647.

Bollen, K. and Curran, P. (2005). *Latent Curve Models*. New York: Wiley Interscience.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Coleman, J. (1968). The mathematical study of change. In H. Blalock and A. Blalock (eds), *Methodology in Social Research*, pp. 428–478. New York: McGraw Hill.

Dorman, C. (2001). Modeling unmeasured third variables in longitudinal studies. *Structural Equation Modeling*, 8: 575–598.

Duncan, O. D. (1975). *An Introduction to Structural Equation Models*. New York: Academic Press.

Finkel, S. E. (1995). *Causal Analysis with Panel Data*. Thousand Oaks, CA: Sage.

Finkel, S. E. (2003). The Impact of the Kenya National Civic Education Programme on Democratic Attitudes, Knowledge, Values and Behavior. United States Agency for International Development Contract AEP-I-00-00-00018-00. Washington, DC: Management Systems International.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge, UK/New York: Cambridge University Press.

Green, D. and Palmquist, B. (1990). Of artifacts and partisan instability. *American Journal of Political Science*, 34: 872–902.

Greene, W. H. (2003). *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology*, 30: 507–544.

Hardin, J. and Hilbe, J. (2003). *Generalized Estimating Equations*. Boca Raton, Fl: Chapman & Hall.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, 46: 1251–1271.

Hausman, J. and Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, 49: 1377–1398.

Heise, D. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review,* 334: 93–101.

Hsiao, C. (2002). *Analysis of Panel Data*. Cambridge UK: Cambridge University Press.

Hu, L. and Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (ed.), *Structural Equation Modeling: Concepts, Issues and Applications*. Thousand Oaks, CA: Sage.

Jagodzinski, W., Kuehnel, S. and Schmidt, P. (1987). Is there a “Socratic effect” in nonexperimental panel studies?, *Sociological Methods and Research*, 15: 259–303.

Jöreskog, K. and Sörbom, D. (1994). *LISREL 8 and PRELIS 8: Comprehensive Analysis of Linear Relationships in Multivariate Data*. Hillsdale, NJ: Erlbaum.

Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions*. Thousand Oaks, CA: Sage.

Kenny, D. A. (1979). *Correlation and Causality*. New York: Wiley.

Kessler, R. C. and Greenberg, D. F. (1981). *Linear Panel Analysis: Models of Quantitative Change*. New York: Academic Press.

Liker, J. K., Augustyniak, S. and Duncan, G. J. (1985). Panel data and models of change: A comparison of first-difference and conventional two-wave models. *Social Science Research*, 14: 80–101.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, 22: 376–398.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49: 1417–1426.

Plümper, T. and Troeger, V. (forthcoming). Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis*

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Singer, J. D. and Willett, J. B. (2002). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, UK/New York: Oxford University Press.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, Fl: Chapman & Hall/CRC.

Sörbom, D. (1982). Structural equation models with structured means. In K. Jöreskog and H. Wold (eds), *Systems under Indirect Observation, Part I*, pp. 183–195. Amsterdam: North Holland.

Teachman, J., Duncan, G., Yeung, W. J. and Levy, D. (2001). Covariance structure models for fixed and random effects. *Sociological Methods and Research*, 30: 271–288.

Tuma, N. B. and Hannan, M. T. (1994). *Social Dynamics*. New York: Academic Press.

Wawro, G. (2002). Estimating dynamic panel models in political science. *Political Analysis*, 10: 25–48.

Wheaton, B., Muthén, B., Alwin, D. and Summers, G. (1977). Assessing reliability and stability in panel models. In D. Heise (ed.), *Sociological Methodology*, pp. 84–136. San Francisco: Jossey-Bass.

Wiley, D. E. and Wiley, J. A. (1970). The estimation of measurement error in panel data. *American Sociological Review*, 35: 112–117.

Woolridge, J. (2002). *Econometric Analysis of Cross-Sectional and Panel Data*. Cambridge, MA: MIT Press.

Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 45: 470–490.

Chapter 30

Panel analysis with logistic regression

Scott Menard

There is an extensive and well-developed literature on panel analysis with continuous or quantitative variables, measured on an interval or ratio scale, with models for panel data that include approaches based on pooling longitudinal and cross-sectional data, fixed effects and random effects models, and structural equation modeling; see, for example, Bijleveld et al. (1998), Finkel (1995), Hardin and Hilbe (2003), Hsiao (2003), Kessler and Greenberg (1981), Maruyama (1998), Sayrs (1989), and Wooldredge (2002), plus the chapters in this volume by Worrall (Chapter 15), Hilbe and Hardin (Chapter 28), Greenberg (Chapter 17), and Finkel (Chapter 29). Less extensively covered, but not entirely absent from this literature, is the use of panel analysis when the dependent variable is a categorical/qualitative dichotomous, nominal, or ordinal variable. Special problems arise in the use of categorical dependent variables in panel analysis, including issues of how to measure change and how to model the within-cases dependency in multiwave panel data. One approach to modeling categorical panel data is to use logistic regression analysis (Hosmer and Lemeshow, 2000; Menard, 2002a). The logistic regression family of models is readily adaptable to both dichotomous and polytomous (nominal or ordinal, and with three or more categories) data, allows the use of standardized coefficients (Menard, 2004a) to compare the magnitude of effects of variables which do not have a natural metric, and is readily accessible in general-purpose statistical software packages such as SAS, SPSS, and Stata.

This chapter provides an applied approach to the use of logistic regression analysis to analyze categorical panel data, beginning with the issue of measuring change, then considering simple two-wave panel models, and ending with a consideration of the use of logistic regression analysis for multiwave panel data. The data used for the examples are taken from the National Youth Survey, a multiwave longitudinal study of a self-weighting national household probability sample of 1725 individuals who were 11–17 years old when first interviewed in 1977, and who were last interviewed in 2002. The dependent variable is marijuana use, or more specifically change in the prevalence (yes or no) of marijuana use. The predictors are (1) exposure to delinquent friends, a scale indicating how many of one's friends have engaged in nine different types of delinquent behavior ranging from assault to theft to illicit and underage substance use and drug sales, plus whether they have encouraged the respondent to do anything against the law; (2) belief that it is wrong to violate the law, a scale indicating how wrong the respondent thinks it is to engage in any of nine types of behavior (the same as the first nine items in the exposure to delinquency scale); (3) age, measured as years since birth; (4) gender, coded 0 = female, 1 = male; and (5) ethnicity, with white non-Hispanic respondents as the reference category and two other categories, African-American and Other. Theoretically, marijuana use should be positively associated with exposure to delinquent friends and negatively associated with belief that it is wrong to violate the law. Age, ethnicity, and gender are included as demographic controls, but associations of age, gender, and ethnicity with marijuana use have been found in past research. For a more detailed description of the sample, the variables, and the theoretical basis for the models tested here, see Elliott et al. (1989).

Let us begin by assuming that we have a dependent or outcome variable of interest, Y, with independent variables or predictors X_k where $k = 1, \ldots, K$, and Y and the X_k are measured at time periods $t = 1, \ldots, T$. (The discussion here does not depend on whether we are concerned with a model that is merely predictive, or whether we want to make causal inferences). As described in more detail in Menard (2002b), there are four "pure" types of causal or predictive models, here presented in bivariate form involving one predictor X and one outcome Y: (A) $X \longrightarrow Y$, where the value of the dependent variable is expressed as a function of the value of the independent variable; (B) $\Delta X \longrightarrow Y$, where ΔX represents a change in X, and the value of the dependent variable is expressed as a function of the change in the independent variable; (C) $X \longrightarrow \Delta Y$, where ΔY represents a change in Y and the change in the dependent variable is expressed as a function of the value of the independent variable; and (D) $\Delta X \longrightarrow \Delta Y$, where the change in the dependent variable is expressed as a function of the change in the independent variable. Models in which the independent variables include both level and rate-of-change variables (e.g., population density and population growth rate as influences on economic development) are also possible. Model (A) requires only cross-sectional data (possibly time-ordered cross-sectional data; see Menard, 2002a, or the brief discussion in Chapter 1 in this volume); the other three models require measurement of at least one of the variables in the model for at least two different times, and models (C) and (D) specifically require repeated measurement of Y. In the simplest instance for longitudinal data analysis, $T = 2$ and we have a two-wave panel model, the case on which we will focus in this chapter.

1 Measuring change in categorical dependent variables

Implicit in the notation of ΔY in the forgoing discussion is the assumption that we know how to measure continuity and change in the dependent variable. For the interval and ratio variables that are used as the dependent variables in linear regression, the options for measuring change are straightforward, involving subtraction of the earlier value of Y from the later value of Y, where Y is measured on an interval or ratio scale. Measuring change for qualitative variables requires more careful consideration of what types of change (or continuity) are of interest. These possibilities can be represented in a contingency table, as illustrated in Figure 30.1. Alternatively, using the percentages for the rows or columns representing the earlier measurement of Y, we can represent the possible patterns of continuity or change in a *transition matrix* which indicates the percentage of cases in each initial category of Y that either remain in that category or switch categories. Figure 30.1 illustrates contingency tables and transition matrices for hypothetical dichotomous and polytomous variables.

Let us assume that we have a dichotomous dependent variable for which 0 = failure and 1 = success. Part A of Figure 30.1 presents a hypothetical contingency table in which we have 500 successes and 500 failures at time 1, and 400 successes and 600 failures at time 2.

(A) Contingency table, dichotomous variable

Frequencies	Time 2 Failure	Time 2 Success
Time 1 Success	200	300
Time 1 Failure	400	100

(B) Transition matrix, dichotomous variable

Row percentages	Time 2 Failure	Time 2 Success
Time 1 Success	.4000	.6000
Time 1 Failure	.8000	.2000

(C) Contingency table, polytomous variable

Frequencies	1	2	3	4
1	200	100	75	25
2	100	175	75	50
3	50	75	150	125
4	25	50	125	200

(D) Transition matrix, polytomous variable

Row percentages	1	2	3	4
1	.5000	.2500	.1875	.0625
2	.2500	.4375	.1875	.1250
3	.1250	.1875	.3750	.3125
4	.0625	.1250	.3125	.5000

Figure 30.1 Contingency tables and transition matrices

Part B of Figure 30.1 presents the corresponding transition matrix, indicating that 60% of the time 1 successes and 20% of the time 1 failures were successes at time 2. Now the question arises whether we are interested in each of the four cells of parts A and B of Figure 30.1 separately, or whether we are willing to combine some of the cells, in operationalizing ΔY.

(a) We can consider each cell separately: (1) continuity of failure in the lower-left cell, (2) continuity of success in the upper-right cell, (3) change from success to failure in the upper-left cell, (4) change from failure to success in the lower-right cell. This gives us a four-category polytomous nominal dependent variable ΔY, which can be analyzed using polytomous nominal logistic regression.

(b) We can decide that we are interested in the differences in types of changes, and in the difference between continuity and each of the different types of change, but not in the difference between continued success and continued failure. This can be accomplished by subtracting Y at time 1, abbreviated Y_{t1}, from Y at time 2, abbreviated Y_{t2}, to obtain a polytomous ordinal dependent variable ΔY coded 0 for continuity, –1 for a change from success to failure, and +1 for a change from failure to success, which can be analyzed using polytomous ordinal logistic regression.

(c) We can decide that we are interested only in whether a change occurs, and not in whether the change is from success to failure or from failure to success. This gives us a dichotomous dependent variable ΔY, which can be coded 0 for no change and 1 for change. This model can be analyzed using dichotomous logistic regression.

(d) We can decide that we are interested only in whether one particular change, for example the change from failure to success, occurred. This also results in a dichotomous dependent variable ΔY, coded 0 if the change did not occur and 1 if the change did occur. Two alternatives within this coding scheme are (1) to include all cases, in which instance cases in all except the bottom-right cell (time 1 failure, time 2 success) would be coded zero, and cases in the bottom-right cell would be coded 1; or (2) exclude those cases in the first row, because they could not possibly change from failure to success (they are all classified as successes in time 1), and consider only the second row, coding cases as 0 for failure and 1 for success at time 2. This approach parallels event history analysis insofar as it defines a set of cases "at risk" of change in the direction of interest and excludes all other cases from the analysis. This second option, however, effectively changes the dependent variable from ΔY to simply Y_{t2}. In either case, the dependent variable can be analyzed using dichotomous logistic regression.

With a polytomous dependent variable, the options are similar, but with a potentially larger number of categories in the dependent variable ΔY.

(a) We can consider each different possibility for continuity and change separately. If the number of categories in Y is c, then we have c^2 cells in the contingency table and hence c^2 possible values for ΔY.

(b) We can ignore differences among continuities and consider only differences among the possible changes, resulting in $c(c-1)$ possible values for ΔY.

(c) For ordered but not nominal polytomous dependent variables, we can consider only whether there is no change, coded 0, a change from a higher to a lower category, coded -1, or a change from a lower to a higher category, coded $+1$, producing a polytomous ordinal dependent variable ΔY (the same as option b for a dichotomous dependent variable).

(d) We can consider only whether a change occurred or not, regardless of which specific change occurred, resulting in a dichotomous dependent variable ΔY which might be coded 0 for no change and 1 for change (the same as option c for a dichotomous dependent variable).

(e) For ordered but not nominal polytomous dependent variables, we can subtract Y_{t1} from Y_{t2}, treating the differences in ranks the same regardless of the initial rank, for example treating movement from category 1 to category 3 as being equal to the movement from category 2 to category 4. This effectively assumes an interval rather than an ordinal scale, and raises the question of whether polytomous ordinal logistic regression or some other technique such as ordinary least squares (OLS) linear regression is most applicable.

What is different about measuring change for categorical variables is that the measure of change is not defined in terms of some mathematical operation or sequence of operations; instead, it is a process of deciding which changes are really different from which other changes, and really coding rather than calculating the dependent variable to produce a measure consistent with our substantive concerns. Even more than in the case of quantitative dependent variables, the question raised by Cronbach and Furby (1970) becomes pertinent: how should we measure change—or should we?

2 Logistic regression for conditional and unconditional change in two-wave panel models

Models involving ΔY as the dependent variable are called *unconditional change models* (e.g., Finkel, 1995). The *conditional change model* has Y_{t2} instead of ΔY as a dependent variable, but includes Y_{t1} as a predictor in the model. Assuming for the moment that none of the predictors is measured as a change score ΔX, the unconditional change model may be written $\Delta Y = (Y_{t2} - Y_{t1}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. The conditional change model may be written $Y_{t2} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \gamma Y_{t1,}$

or alternatively $(Y_{t2} - \gamma Y_{t1}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. Compared to the logistic regression model for cross-sectional data, the conditional change model just adds one more predictor, the lagged dependent variable $Y_{t1.}$ Comparing the equations for the unconditional and conditional change models, the unconditional change model effectively assumes that $\gamma = 1$, regardless of the other predictors in the model (hence unconditionally), while in the conditional change model the value of γ is conditional on (controls for) the other predictors in the model. Also, as described above, the unconditional change model for categorical dependent variables raises the issue of how to measure change in the dependent variable. In the conditional change model, how to measure ΔY need not be an issue. Instead, the usual procedures for estimating dichotomous, nominal polytomous, or ordinal polytomous logistic regression models may be used, the only change being that the lagged value of the dependent variable (Y_{t1}) is included in the model. (As discussed below, the measurement of Y_{t1} and Y_{t2} may not be identical for a polytomous nominal dependent variable, but this is at worst a minor issue.) Finally, both the unconditional and the conditional change models can be extended to include more than two time periods.

The coefficient of the lagged endogenous variable, γ, is sometimes called the “stability coefficient”. There are several interpretations of the stability coefficient which are statistically indistinguishable; which interpretation is most appropriate must be decided based on conceptual or theoretical considerations (Davies, 1994; Finkel, 1995; Kessler and Greenberg, 1981; Rogosa, 1995). Most commonly, it is interpreted either as a control for prior, unmeasured influences on Y, or as the inertial effect of past values of Y on the present value of Y. Alternatively, it may be interpreted as doing several things at once. Davies (1994) indicates that the stability coefficient may represent the impact of a previous state or behavior on a present state or behavior, plus prior impacts of measured variables, plus effects of unmeasured variables, on the dependent variable Y_t. Because the stability coefficient may be incorporating more than a single type of effect, the conditional change model typically provides a liberal estimate of inertial effects, and a conservative estimate of the effects of other predictors in the model. In this respect, as Davies (1994, pp. 36–37) notes, the conditional change model is far from perfect; but in light of the observation by the same author (Davies, 1994, p. 32) that the impact of interventions is often less than that predicted by statistical models, this characteristic may actually be a desirable feature of the conditional change model.

In both the conditional and unconditional change models, the predictors may or may not be cast as change scores, either X or ΔX. If, in the unconditional change model, all of the predictors are also change scores, ΔX, so change scores are used for the predictors as well as the dependent variable, we have a *first difference* model, model D above. The first difference model is so named because it takes the difference between adjacent measurement periods (the first difference, in time series terminology), but not differences between first differences (the second difference) or higher ordered differences. In first difference and higher order difference models, in addition to the usual issues involved in change models, predictors that do not change from one period to the next (stable individual characteristics) are eliminated from the model, potentially, as noted by King (1989) in the context of time series models, leaving more variation attributable to error. The first difference form of the unconditional change model may be estimated by either ordinary dichotomous logistic regression, or (as detailed below), if we are interested in adjusting for individual differences in patterns of change, conditional logistic regression. (The paradox in terminology, using conditional logistic regression to estimate an unconditional change model, simply reflects the different things being conditioned in the change model and the estimation technique.)

A consideration in both the conditional and unconditional change models is when to measure the predictors relative to the dependent variable. Measuring the predictors and the dependent variable for the same time introduces ambiguity in the time ordering of cause and effect, and may produce overestimates of the impact of the predictor on the dependent variable when there are spurious time-specific effects. Measuring the predictors for a time prior to the time for which the dependent variable is measured helps eliminate this ambiguity in temporal ordering (but does not remove it completely; see, e.g., Menard, 2002b, pp. 18–21), but raises the possibility that the time lag between cause and effect may be too long, resulting in an underestimation of the impact of the predictor on the dependent variable.

In the context of linear regression models with interval or ratio-scaled dependent variables, there has been considerable disagreement in the social and behavioral sciences about the relative appropriateness of unconditional as opposed to conditional change models when the purpose is to analyze change in panels with a small number of periods. This debate is relevant primarily to (a) the analysis of short-term change *within* individuals, (b) when the dependent variable is measured as $\Delta Y = Y_{t2} - Y_{t1}$ on an interval or ratio scale. Arguments against the use of the unconditional change model are that change scores ΔY are systematically related to any random error of measurement, that they are typically less reliable than the raw scores of the variables from which they are calculated, and that the unreliability of change scores may lead to fallacious conclusions or false inferences (Cronbach and Furby, 1970). Arguments in favor of the use of change scores, usually cast in the context of the unconditional change model, are based on the assumption that we are interested in explaining intraindividual change rather than in causal analysis of differences among individuals; that the number of periods is small (typically no more than three); and that certain other assumptions are met. Liker et al. (1985) demonstrated that the unconditional change model may be superior to both cross-sectional equations and the conditional change model when (a) regression parameters remain constant from one period to another, (b) there are unmeasured variables that influence the dependent variable but do not change over time, (c) there is autocorrelated error in the measurement of those variables which both influence the dependent variable and vary over time, and (d) the panel data give more reliable measurement of *changes* in predictor variables over time than of the level or value of predictor variables at any given time, as may be the case if interindividual differences in change are large relative to interindividual differences in initial scores.

The conditions under which the unconditional change model is preferable to the lagged endogenous variable model are quite restrictive, unlikely to be met in most observational research, and difficult to meet even in experimental or quasi-experimental research (Bijleveld et al., 1998; Cronbach and Furby, 1970; Finkel, 1995). In addition, the conditional change model may be more appropriate when there is a true causal effect of Y_{t1} on Y_{t2}. As Davies (1994, p. 33) explains, “positive temporal dependence, or inertia, is to be expected of most social behaviour.” Finkel (1995, p. 7) notes that the prior value of Y may influence the current value of Y, and the influence of Y_{t1} on Y_{t2} may be misspecified by an unconditional change model. On balance, it appears that the conditional change model is more generally valid than the unconditional change model. In addition, in the present context of qualitative dependent variables, it is easier to decide what model to estimate for the conditional change model. As noted earlier, there are several possibilities for operationalizing the measurement of ΔY for a categorical variable in the unconditional change model, each of which requires a different method of calculation (or coding) and different interpretation of the results. In the conditional change model, however, the dependent variable as modeled can often have the same coding as the dependent variable as originally coded, the approach to estimating and interpreting the model is simply the usual procedure for dichotomous or polytomous logistic regression analysis, and the model uses at most c-1 (where c is the number of categories in the dependent variable) sets of coefficients for predicting assignment to the different categories of a polytomous dependent variable (and possibly only one set of coefficients for the predictors in either dichotomous logistic regression or an equal slopes/unequal intercepts ordinal logistic regression model for a polytomous ordinal dependent variable).

3 The subject-specific model for a two-wave panel

In both the conditional and unconditional change models, each case (at time 1) serves as its own control (for time 2), but in the unconditional change model, the strength of the relationship between the scores at time 1 and time 2 is not considered to be a parameter we are interested in estimating. An implicit assumption in both of these models is that the change process is the same across individuals. In particular, it is assumed that the impact of the independent variables on the outcome, as measured by the β parameters, is the same for each case, and that each case has the same marginal probability, as measured by α. As described by Agresti (2002), this model is called the *marginal* model, because it focuses on the marginal distributions of responses for the observations, and the effects in the marginal model are termed *population averaged* effects, because the effects are all averaged over the entire population or sample, rather than being measured separately for each case.

An alternative assumption is that the process of change is unique for each individual, and that we must somehow incorporate this uniqueness into the analysis. In this conceptualization, each case is represented by (at least) two observations. The extreme situation would be one in which each parameter varied by individual, producing the model $\text{logit}(Y) = \alpha_i + \beta_{1i}X_1 + \beta_{2i}X_2 + \ldots + \beta_{Ki}X_K$, where the subscript $i = 1, 2, \ldots, n$ refers to the specific case, and the first subscript on the β coefficients corresponds to the subscript for the K predictors $X_1, X_2, \ldots, X_K$. With only two observations per case, one cannot realistically estimate this model. A simpler model assumes that the effect parameters are constant across cases, but that the intercepts vary by subject, producing the model $\text{logit}(Y) = \alpha_i + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_KX_K$, which differs from the marginal model only in the subscript on the intercept, indicating a different intercept for each case. According to Agresti (2002, pp. 414–415), this model may be described as a *conditional* model, "since the effect β is defined conditional on the subject. . . . The effect is *subject-specific,* since it is defined at the subject level." Actually, Agresti was referring to a model with only one predictor, and it is only the intercept that is subject-specific. With more observations per case, it is possible to construct models in which the β parameters as well as the intercept are subject-specific. This model essentially differs from the marginal model by allowing each case to have its own probability distribution, as represented by α_i. In this model, if the α coefficients are large relative to the β coefficients, the (shared) α_i within each case may determine the outcome, resulting in the same value of Y for both time periods. Alternatively, if α_i is small relative to the β parameters, it is possible to have different outcomes at the two time periods or for the two matched subjects. In this sense, the outcome for one of the paired observations is not necessarily independent of the outcome on the other, and this dependence needs to be taken into account unless all of the α_i are equal.

The large number of parameters α_i is problematic for fitting the model using maximum likelihood, whose fit depends on the number of cases being large relative to the number of parameters. In the marginal model, the number of cases can increase indefinitely while the number of parameters remains small, thus resulting in the large sample properties that make maximum likelihood estimation advantageous. When, instead, with the addition of every new case a new parameter is added to the model as well, the maximum likelihood estimates will be inconsistent (Andersen, 1970; Chamberlain, 1980; Neyman and Scott, 1948). The problem can be resolved if the α_i are treated as "nuisance parameters" in which we have no direct interest. This is done in *conditional logistic regression*. Instead of using a likelihood function that explicitly includes the α_i, conditional logistic regression "conditions" the likelihood function on *sufficient statistics* for the α_i.

A sufficient statistic summarizes the information about a particular population parameter (such as the α_i) such that if some function of the outcome Y depends on the distribution of the parameter, the sufficient statistic incorporates sufficient information about that parameter that the conditional distribution of the sample depends on the sufficient statistic and not, in addition to the sufficient statistic, the parameter itself. In conditional logistic regression, the sufficient statistics for the α_i are their pairwise totals of "successes". If we set $Y_{i1} = 0$ for "failure" and $Y_{i1} = 1$ for "success" at time 1 for case i, and $Y_{i2} = 0$ for "failure" and $Y_{i2} = 1$ for "success" at time 2 for the same case i, then the sum $S_i = Y_{i1} + Y_{i2}$ is the pairwise total of successes for Y_i. When the observations are identical on Y, the sufficient statistic S_i is equal to either 2 (both successes) or 0 (both failures); when $S_i = 1$, the outcomes differ. As described by Agresti (2002, p. 416), the distribution of (Y_{i1}, Y_{i2}) depends on $\boldsymbol{\beta}$ only when the outcomes differ for the two responses, and the conditional distribution is equal to $\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K)]/[1 + \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K)]$ when $Y_{i1} = 0$ and $Y_{i2} = 1$, and equal to $1/[1 + \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K)]$ when $Y_{i1} = 1$ and $Y_{i2} = 0$, where "exp(Y)" represents the exponential function e^Y.

4 Estimation of the conditional logistic regression model

Let $i = 1, 2, \ldots, n$ denote the cases. Within each case, let $t = 1, 2, \ldots, T_i$ denote the observations. Let the dependent variable Y take on observed values $y_{i1}, y_{i2}, \ldots, y_{iT}$, where each y_{it} is equal to either zero or one, for the T observations within each case. Let there be K predictors, $X_1, X_2, \ldots, X_K$. Let h_{1i} be the number of cases for which $y_{it} = 1$, or equivalently the sum (over t) of the observed values $\Sigma_t y_{it}$. In the most general instance, the number of cases for which $y_{it} = 1$ is h_{1i} and the number of cases for which $y_{it} = 0$ is $h_{2i} = T_i - h_{1i}$ within each case. Using boldface type here to designate vectors, let $\mathbf{x_{it}}$ represent the vector of values of the predictors $X_1, X_2, \ldots, X_K$, let $\boldsymbol{\beta}$ represent the vector of K coefficients for the predictors, $\beta_1, \beta_2, \ldots, \beta_k$, and let exp(Y) represent the exponential function e^Y. Let S_i be the set of all possible combinations of h_{1i} ones and h_{2i} zeros in case i, and let d_i be an element of S_i. The individual components of each (vector) element of S_i are designated d_{it}, and correspond to the possible values for each corresponding y_{it} for different possible vectors $\mathbf{y_i}$.

Given these definitions, the probability of a possible set of values of Y for an individual, the vector $\mathbf{y_i} = \{y_{i1}, y_{i2}, \ldots, y_{iT}\}$, is equal to $P[\mathbf{y_i}|\Sigma_t y_{it} = h_{1i}] = \exp(\Sigma_t y_{it}\mathbf{x_{it}}\boldsymbol{\beta})/[\Sigma_{di}\exp(\Sigma_t d_{it}\mathbf{x_{it}}\boldsymbol{\beta})]$. That is, the probability of a particular set of *observed* values for Y for a particular case is equal to the exponentiated sum (over all observations in the case) of the products involving the *observed* values of Y, the observed values of the predictors, and the coefficients of the predictors, divided by the exponentiated sum (over all possible combinations of values for Y in the case) of products of the *possible* values of Y (d_{it}), the observed values of the predictors, and the coefficients of the predictors. The corresponding conditional log-likelihood is $L = \Sigma_i\{\Sigma_t(y_{it}\mathbf{x_{it}\beta}) - \ln[\Sigma_{di}\exp(\Sigma_t d_{it}\mathbf{x}_{it}\boldsymbol{\beta})]\}$. This conditional probability does not involve the α_i, so the α_i are not estimated when the conditional likelihood is used.

Table 30.1 Conditional logistic regression fixed effects model

Dependent variable	$R_L{}^2/R_O{}^2$	*Independent variables*	*b**	*b*	*p(b) (Wald)*
Change in marijuana use	0.016	Exposure	−0.128	−0.029	0.403
	0.133	Belief	−0.421	−0.099	0.027

5 The fixed effects model using conditional logistic regression

In the conditional logistic regression model, as described above, $\text{logit}(\Delta Y) = \Delta\alpha_i + \beta_1\Delta X_{1i} + \beta_2\Delta X_{2i} + \ldots + \beta_K\Delta X_{Ki}$, where i indexes the cases within which the observations are clustered. The unit-specific intercepts are assumed to be constant within the cases, so $\Delta\alpha_i = 0$ and drops out of the equation. Estimating this model using full maximum likelihood results in inconsistent estimates for the α_i and the β_k (Andersen, 1970; Chamberlain, 1980). Although the conditional logistic regression model most closely parallels model D at the beginning of this chapter (the unconditional change model with change scores as predictors, $\Delta X \longrightarrow \Delta Y$), it is different from the model obtained using unconditional logistic regression in several respects: (1) cases for which Y is a constant (and hence $\Delta Y = 0$) are dropped from the analysis, thus potentially excluding cases for which the predictors have consistently high or consistently low values within the case over time; (2) predictors which were constant within cases were also dropped from the model, thus potentially excluding predictors which might, although constant, affect the probability of change; (3) it includes the assumption of fixed within-case effects, α_i, and if the within-case effect is not fixed, then the model is misspecified; and (4) the model is actually "retrodicting" the independent variables from the dependent variable. Regarding this last point, it is actually the y_i that are fixed (only those cases for which ΔY=1 are included in the analysis), and it is the values of the $\mathbf{x}_{it}$ that are allowed to vary. All of these differences may lead to results which may be different from results produced using other methods of estimation for longitudinal models.

In Table 30.1, a conditional logistic regression model is presented for the unconditional change in marijuana use from time 1 to time 2 using data from the National Youth Survey (Elliott et al., 1989) for the years 1979 and 1980. The theoretical predictors for change in marijuana use are exposure to delinquent friends, belief that it is wrong to violate the law, age, gender, and ethnicity (White, African-American, Other) as predictors. In principle, conditional logistic regression is not limited in the number of time periods that can be incorporated into the model, but use of a large number of time periods makes it important to consider nonlinear relationships between the dependent variable and the time dimension, if the time dimension (age or period) is included in the model.

The first column in Table 30.1 lists the dependent variable. The second column lists two measures of explained variation, the likelihood ratio coefficient of determination $R_L{}^2$ and the squared correlation between the observed (0 or 1) and predicted (continuous probability between zero and one) values of the dependent variable, $R_O{}^2$. Arguments have been made for the latter measure in the context of logistic regression analysis, but evidence to date appears to indicate that for logistic regression analysis using maximum likelihood estimation, $R_L{}^2$ is the more appropriate measure of explained variation, although the two typically produce similar results (Menard, 2000). The third column lists the predictors, and the fourth column lists the fully standardized logistic regression coefficients, which are directly analogous and behave similarly to the standardized coefficients used in OLS regression (Menard, 2004a). The fifth column lists the unstandardized logistic regression coefficients, and the last column lists their statistical significance based on the Wald test. For a general treatment of standardized and unstandardized coefficients in logistic regression analysis, see Menard (2002a).

In Table 30.1, only 462 observations were used in calculating the conditional logistic regression model. At 2 observations per case, this means only 231 of the 1725 cases had valid data indicating change, with most (1257) having had valid data indicating *no* change in the dependent variable. As noted earlier, cases for which there is no change are dropped from the analysis, and this example illustrates the loss of data that can occur when using conditional logistic regression to estimate the unconditional change model accounting for interindividual heterogeneity (the individual-specific intercepts α_i). In addition, the only predictors in the first row of Table 30.1 are Exposure and Belief; gender and ethnicity are constants within cases, and so because they do not vary over time they are necessarily dropped from the model, and the difference in age is a constant (everyone is one year older in the later wave than in the earlier wave), so age, too, is excluded from the model.

The influence of exposure is unexpectedly in the negative direction, but not statistically significant; the influence of belief is, as expected, negative and statistically significant, with a standardized coefficient of $-.421$. $R_L{}^2 = .02$, only marginally significant at $p=.0751$, but $R_O{}^2 = .13$, suggesting a slightly stronger relationship (and a larger discrepancy than one would normally expect in the conclusions to be derived from the two measures of explained variation). These results are distinctly at odds with other findings (e.g., Elliott et al., 1989) that exposure is the strongest predictor of marijuana use. Why the difference? Part of the explanation lies in the fact that individuals with very high levels of exposure who use marijuana at both times, and individuals with very low levels of exposure who use marijuana at neither time, have been eliminated from the analysis because the dependent variable is a constant. These results, then, suggest caution when applying and interpreting the fixed effects model using conditional logistic regression to the analysis of longitudinal data. The exclusion of cases which do not change on the dependent variable is potentially highly problematic, as these may be the cases which contribute the most to the association between the dependent variable and the predictors.

6 Unconditional logistic regression for the unconditional change model

Using unconditional as opposed to conditional logistic regression for the unconditional change model, four models were calculated for change in marijuana use as the dependent variable. In all four models, change in marijuana use is measured by subtracting marijuana use in 1979 from marijuana use in 1980, resulting in an ordinal dependent variable coded -1 for change from use to nonuse, $+1$ for change from nonuse to use, and 0 for no change (both users and nonusers). In the first model, constructed to parallel as closely as possible the conditional logistic regression analysis above, change in exposure and change in belief are the predictors, calculated as 1980 exposure or belief minus 1979 exposure or belief. Here, the choice is made to consider changes in the independent variables that occur contemporaneously with the change in the dependent variable. The alternative, measuring change in the independent variables strictly prior to change in the dependent variable (i.e., measuring change in the predictors from 1978 to 1979), would result in a model in which the time ordering was clearly correct between predictor and outcome, but in which the lag time might be too long, resulting in lower levels of explained variation and smaller standardized and unstandardized coefficients for the predictors. In the second model, age is added to the equation. In the third model, age is dropped and gender and ethnicity are added to the equation. The fourth and final analysis includes age, gender, ethnicity, change in exposure, and change in belief as predictors. Analysis is performed using the cumulative logit model, and the results are presented in Table 30.2. The threshold coefficients in Table 30.2 are analogous to intercepts in OLS regression and dichotomous logistic regression, and are of no substantive interest, but are presented for completeness. Note that we have now moved from a fixed effects to a marginal model.

In Table 30.2, each successive model explains a little more of the variation in the dependent

Table 30.2 Unconditional change models for marijuana use

Dependent variable	$R_L{}^2/R_O{}^2$	*Independent variables*	b^*	b	*p(b) (Wald)*
Change in marijuana use	.026	Change in exposure	.146	.114	.000
	.028	Change in belief	−.057	−.051	.030
		Threshold 1	–	−2.798	.000
		Threshold 2	–	2.306	.000
	.033	Change in exposure	.142	.113	.000
	.034	Change in belief	−.054	−.048	.039
		Age	−.082	−.082	.002
		Threshold 1	–	−4.788	.000
		Threshold 2	–	.361	.566
	.034	Change in exposure	.155	.120	.000
	.036	Change in belief	−.058	−.051	.029
		Gender (male)	.005	.026	.855
		Ethnicity:			
		African-American	.087	.656	.001
		Other	.033	.381	.214
		Threshold 1	–	−3.750	.000
		Threshold 2	–	1.401	.000
	.039	Change in exposure	.151	.119	.000
	.042	Change in belief	−.054	−.049	.038
		Age	−.080	−.113	.003
		Gender (male)	.008	.045	.755
		Ethnicity:			
		African-American	.083	.638	.001
		Other	.027	.319	.299
		Threshold 1	–	−5.600	.000
		Threshold 2	–	−.406	.565

variable than the previous model, although the overall level of explanation is low, according to both R_L^2 and R_O^2. This appears to be because the cases are heavily clustered in the "no change" category of the dependent variable. For all of the models, the standard test for nonparallel slopes in ordinal logistic regression analysis was not statistically significant, indicating that the parallel slopes model fits well. Turning to the individual predictors, change in exposure is consistently statistically significant and the strongest predictor, based on the standardized coefficient. An increase in exposure is associated with higher change scores for marijuana use. When they are included in the model, age and African-American ethnicity are also statistically significant and nearly equal in magnitude. Increasing age is negatively associated, and African-American ethnicity is positively associated with the change score for marijuana use. Further exploration of these relationships suggest that (a) marijuana use tends to be initiated at younger ages, and to decline at older ages, producing the pattern of results with age, and (b) African-American respondents tend to initiate marijuana use later than non-African-American respondents, with the result that they are more likely than non-African-American respondents to be increasing their marijuana use at the ages (14–21) covered in these waves of the NYS. Fourth in magnitude of effect is change in belief, with the expected negative relationship between stronger beliefs that it is wrong to violate the law and higher change scores for marijuana use. Gender and "Other" ethnicity are not statistically significant predictors of marijuana use. Males are no more likely than females, and "Other" ethnic groups no more likely than white non-Hispanic Europeans, to have higher change scores on marijuana use.

A few comments on these results are in order. First, the relatively low levels of explained variation when using change scores (for exposure and belief) as predictors of change scores are not uncommon in social science research. Second, the use of change scores as predictors often introduces one of two problems, excessively long lag between change in independent variables and change in independent variables, or else ambiguous temporal (and hence causal) ordering between changes in independent variables and changes in the dependent variable. Third, even though African-American ethnicity itself does not change, it does have an impact on change in marijuana use, reinforcing the observation that the exclusion of time-constant variables from the conditional logistic regression model described in the previous section may be problematic. Fourth, aside from the influence of African-American ethnicity on change in marijuana use and the sheer magnitude of the standardized coefficients and explained variation, these results are fairly similar to cross-sectional models for marijuana use using the same predictors and prevalence of marijuana use as a dependent variable. Using raw scores instead of change scores for exposure and belief in the present example (results not shown in detail here) merely produces lower levels of explained variation than those obtained using change scores. The model predicting change in marijuana use from change, rather than level, of exposure and belief appears to be reasonably well specified; it just does not explain a great deal of variation in the change in marijuana use. Fifth, as noted earlier in this chapter, there are other codings that could be used for change in marijuana use; for all of them, the unconditional change model seems best suited to analysis of change from one time period to a subsequent period, that is, to the analysis of only two periods rather than multiple periods. With data from multiple periods, however, one could use a multiwave panel analysis approach parallel to that described by Finkel in Chapter 29, with logistic regression for the unconditional change model, possibly incorporating chronological or calendar time in addition to or instead of age as a control variable.

This possibility will be considered further at the end of this chapter.

Finally, despite their apparent similarities, the models in Table 30.1 and the first row of Table 30.2 are different in a very important way. The model in Table 30.1 assumes (correctly or incorrectly) that (1) there is unmeasured heterogeneity in the respondents, (2) the form taken by this unobserved heterogeneity is adequately reflected in differences in intercepts, representing unmeasured interindividual differences in the distribution of the probability of marijuana use, and not in the coefficients, which would represent unmeasured interindividual differences in the *effects* of exposure and belief on marijuana use; and (3) the exclusion of cases with no change, and predictors which are either constant over time (gender and ethnicity) or whose changes are invariant across individuals (age) does not bias the results. In contrast, the models in Table 30.2 assume either that there is no significant unmeasured heterogeneity among the respondents that would be reflected in either intercept or slope parameters (equal slopes are also assumed in the model in Table 30.1), or that we are not concerned with such individual variation (e.g., because we are planning an intervention at the population rather than the individual level, and it is the average effects across individuals, not the individual-level effects that count). Even given this, the results from the conditional logistic regression fixed effects model render it suspect, given past research in this area. Better resolution of this issue might be possible with additional waves of data, allowing the modeling of individual trajectories (not just intercepts) and perhaps random slope coefficients; that possibility is addressed very briefly at the end of this chapter, but see also Chapter 33 in this volume (Menard) regarding the use of multilevel longitudinal analysis of categorical dependent variables.

7 Logistic regression for the conditional change model

For the conditional change model, four models were calculated, paralleling the models for the unconditional change model, and they are presented in Table 30.3. The dependent variable is current marijuana use in wave 5 (1980) of the NYS. The predictors are exposure (not change in exposure), belief (not change in belief), age, gender, and ethnicity, plus prior marijuana use. In this analysis, exposure and belief are measured at the same time as *prior* marijuana use (1979), temporally prior to the dependent variable, current (1980) marijuana use. The temporal ordering of the presumed cause and effect is thus relatively unambiguous. It would be possible to include change scores for exposure and belief from 1979 to 1980 in place of levels of exposure and belief in 1979; the use of the levels as opposed to the change scores is, however, the more common practice in conditional change models, in part because these models can be interpreted not only as models of change in the dependent variable, but also as models of the level of the dependent variable that simply include the lagged dependent variable (in this instance prior marijuana use) as a predictor.

The first column in Table 30.3 specifies the dependent variable, as in previous tables. The second column includes R_L^2, ΔR_L^2, and ΔR_O^2, where ΔR_L^2 is the change in R_L^2 that occurs when the other predictors are added to a model that already includes prior marijuana use. In effect, ΔR_L^2 is the *likelihood ratio partial multiple squared correlation:* (a) likelihood ratio (using R_L^2), (b) partial (controlling for prior marijuana use), (c) multiple (since it involves two or more predictors), (d) squared correlation (really explained variation) between the set of predictors and current marijuana use, controlling for prior marijuana use. Comparing ΔR_L^2 in Table 30.3 with R_L^2 in Table 30.2, the two are of a similar order of magnitude, with ΔR_L^2 being very slightly larger, a common result

Table 30.3 Conditional change models for marijuana use

Dependent variable	R_L^2/ $\Delta R_L^2/R_O^2$	*Independent variables*	*b**	*b*	*p(b) (Wald)*
Marijuana	.401	Exposure	.144	.088	.000
use	.042	Belief	−.187	−.129	.000
	.482	Prior marijuana use	.402	2.270	.000
		Intercept	–	.960	.192
	.402	Exposure	.143	.087	.000
	.042	Belief	−.194	−.133	.000
	.482	Age	−.025	−.036	.368
		Prior marijuana use	.489	2.749	.000
		Intercept	–	1.657	.121
	.404	Exposure	.146	.088	.000
	.043	Belief	−.199	−.136	.000
	.482	Gender (male)	−.027	−.151	.327
		Ethnicity:			
		African-American	.045	.345	.092
		Other	.017	.195	.536
		Prior marijuana use	.489	2.737	.000
		Intercept	–	1.155	.121
	.404	Exposure	.144	.087	.000
	.044	Belief	−.205	−.140	.000
	.482	Age	−.026	−.037	.358
		Gender (male)	−.028	−.152	.323
		Ethnicity:			
		African-American	.045	.348	.089
		Other	.015	.182	.567
		Prior marijuana use	.495	2.766	.000
		Intercept	–	1.870	.083

in comparing unconditional and conditional change models. As a partial correlation, ΔR_L^2 is a very conservative estimate of the impact of the predictors on the dependent variable.

Turning to the individual predictors, the substantive results are similar but not identical to the results for the unconditional change model. First, prior marijuana use is the strongest influence in each of the models in Table 30.3, followed by belief (with standardized coefficients $b^* = -.187$ to $-.205$), then exposure ($b^* = .143$ to $.146$), and all three of these predictors are statistically significant in all four models. Being African-American is the only other influence to attain even marginal significance, with $b^* = .083$ to $.087$, $p = .089$ to $.092$. In contrast to the results for the unconditional change model, age does not appear to be a statistically significant influence on marijuana use. Note, however, that age was a very weak ($b^* < .100$) influence in the unconditional change model, and would have been regarded as substantively nonsignificant, even though it was statistically significant. The coefficients for exposure are comparable for the conditional and unconditional models, but the conditional model suggests that the influence of belief is substantively as well as statistically significant (and stronger than the influence of exposure), while the unconditional model indicates that the influence of belief is substantively nonsignificant and weaker than the influence not only of exposure but also of being African-American. Bear in mind that, in principle, these differences in results could reflect not only the differences between dependent variables in the conditional and unconditional change model, but also differences in the predictors (levels as opposed to changes in exposure and belief). As indicated earlier, however, the results of the unconditional change model are similar regardless of whether levels or changes in belief and exposure are used as predictors, with slightly lower explained variation using levels as opposed to changes in belief and exposure in the unconditional change model, so it seems more likely that it is not the differences in how the predictors are measured, but rather the differences between the operationalization of change in the dependent variable, that makes the difference between the conditional and unconditional change models here.

8 Extensions to polytomous dependent variables

The conditional and unconditional change models are readily extended to the analysis of polytomous dependent variables. In the unconditional change model for polytomous dependent variables, the question is how to code change in the dependent variable and in any of the time-varying categorical predictors, an issue already addressed above. In the conditional change model, the question is how to code the lagged endogenous variable. For a nominal polytomous dependent variable, the baseline logit model compares each category with a baseline category. This means that duplicating the coding of the dependent variable for the lagged dependent variable results in missing cases for each of the dummy lagged variables, and the model cannot be calculated. Instead, some other contrast will be necessary. One possibility here is to use indicator coding (1 for the category in question, zero for all other categories, not just for the reference category) for the predictors, but other contrasts (e.g., deviation or effect coding; see Menard, 2002a for a more extended discussion of contrasts for categorical predictors in logistic regression analysis) are also possible. For ordinal polytomous dependent variables, whether the coding of the lagged endogenous variable can be identical to the coding of the dependent variable depends on which of the ordinal logistic regression models is specified. In particular, for the cumulative logit model, which uses all of the cases for each of the contrasts, it is possible to code the dummy variables representing the lagged endogenous variable Y_{t-1} identically to the dummy variables representing the dependent variable Y_t. Alternatively, one can treat the lagged endogenous variable Y_{t-1} as either a set of unordered dummy variables with an indicator (or other) contrast, or one can treat Y_{t-1} as an interval predictor, but this latter strategy raises the issue of whether, if Y can be treated as an interval variable, logistic regression is the most appropriate approach to its analysis.

9 Multiwave logistic regression panel models

So far, we have been dealing only with two-wave panel models, models in which measurement of the dependent variable (and possibly the predictors) occurs at two distinct times. It is also possible to have a two-wave panel model in which more than one variable is treated as a dependent variable, at least with respect to some of the other variables in the model. For example, it is possible that marijuana use not only is influenced by, but also influences, exposure to delinquent friends and belief that it is wrong to violate the law. In the terminology of path analysis, belief, exposure, and marijuana use may all be treated as endogenous variables with respect not only to gender and ethnicity, but also with respect to one another, in a two-wave or multiwave panel model. As with any other path analytic model, we must be concerned with method of estimation and model identification. If we build all possible influences into the model, the model will be underidentified, and we will not be able to calculate the path coefficients for the model until we impose some constraints upon the model.

One common constraint is to limit the influence of an independent variable to the wave immediately following the wave for which the independent variable was measured, and to assume that any impact of the independent variable on subsequent waves occurs as a result of, or in a sense is filtered through, the variables in the immediately following wave on which the independent variable has an effect. For example, if exposure has an effect on marijuana use, it may be assumed that exposure at time 1 directly affects marijuana use at time 2, but does not affect marijuana use at time 3, independent of its effect on exposure at time 2 and marijuana use at time 2 (which in turn may affect marijuana use at time 3). Another commonly imposed constraint for a small number of waves (e.g., 1–5) is to impose the constraint that the effect of one variable, for example belief, on another variable, for example exposure, is assumed to be constant regardless of which adjacent waves the variables are measured.

Alternatively, one can allow for the possibility that the impact of an independent variable on a dependent variable varies over time or age, either by building interaction terms between time or age and other variables into the model, or if none of the coefficients is constrained to be equal over time or age, by separate estimation of each equation (as opposed, for example, to using a pooled TSCS approach). Separate estimation of multiple equations in a system of structural equations was more common in the earlier days of path analysis, and has largely fallen out of favor with the development of simultaneous estimation techniques such as LISREL for structural equation models. This is because the use of full-information maximum likelihood techniques allows more efficient use of valid theory, resulting in more precise parameter estimates; but even for structural equation models for quantitative variables, separate estimation does have the advantage that errors in one part of the model are not propagated throughout the rest of the model (Heise, 1975). For logistic regression analysis, the absence of simultaneous equation estimation techniques parallel to those for quantitative dependent variables is an additional reason for considering a separate estimation approach for structural equation or path analysis models using logistic regression (Menard, 2004b).

Conditional change models for dichotomous and ordinal variables can also be implemented in the structural equation modeling framework, using polychoric correlations and weighted least squares (Jöreskog and Sörbom, 1989) as an alternative to the logistic regression framework. Another consideration in two-wave or multiwave panel models with multiple endogenous variables is that we do not have truly independent measures from one wave to the next, but instead have observations that are nested within individuals (and possibly within primary sampling units as well). Knowledge of this potential source of dependency in the data suggests that the use of robust standard error estimates or generalized estimating equations (Hardin and Hilbe, 2003; see also Hilbe and Hardin, Chapter 28, in this volume) should be considered.

With five or more waves of data, other solutions become available. The time dimension can be explicitly incorporated into the model, including nonlinear functions of time, in multilevel analysis using the general linear model (e.g., Raudenbush and Bryk, 2002; see also Menard, Chapter 33, in this volume). Another possibility is the use of discrete time event history analysis (e.g., Singer and Willett, 2003; Yamaguchi, 1991; see also Keiley et al., Chapter 27, in this volume) to model not only whether but also when an outcome occurs. It remains possible, particularly if one is interested in modeling relationships that are conceptually reciprocal (exposure influences marijuana use and marijuana use influences exposure), to use separate estimation of recursive models for each time-specific dependent variable to explore more complex causal structures, parallel to those analyzed using OLS regression (for separate estimation) or maximum likelihood structural equation models (for simultaneous estimation). As noted above, separate estimation has largely fallen out of favor for models with continuous interval/ratio dependent variables, for which maximum likelihood structural equation modeling is now the standard, but simultaneous equation techniques for estimation of multiwave panel models with categorical dependent variables are not well developed or readily available in existing statistical software, and separate rather than simultaneous estimation of equations may be the most practical solution at present.

10 Conclusion

The use of logistic regression in longitudinal panel analysis poses all of the problems associated with the use of ordinary least squares linear regression or related techniques in longitudinal analysis, plus a few more. As with linear regression, there are questions of (1) whether to use a conditional or an unconditional change model, and (2) whether to measure predictors as levels (the predictor measured at a single time) or change scores. In addition, for longitudinal logistic regression models, we must decide how to measure change in qualitative dependent variables, as dichotomies, trichotomies, considering each possible change separately, and whether to consider each possible type of continuity separately. For much the same reasons that the conditional change model is more generally applicable in linear panel analysis, and for other reasons as well, the conditional change model seems generally to be the best option for data involving a large number of cases and relatively few time periods, either dichotomous or polytomous (nominal or ordinal) data with relatively few categories, and the use of logistic regression analysis. The conditional change model is also consistent with approaches taken in analyzing data with a larger number of periods, particularly multilevel change models. This is not to say that one should avoid the use of other models for longitudinal analysis of data with a small number of periods, only that one should probably begin with the conditional change model and consider whether the data, model assumptions, or other concerns provide sufficient reason for selecting a different model for longitudinal panel analysis using logistic regression.

Software

SAS, SPSS, Stata, and other general-purpose statistical software packages include routines for dichotomous and polytomous logistic regression analysis. None is completely satisfactory, and none generates all of the coefficients used in the present chapter without some additional calculation by hand; see Menard (2000, 2002b, 2004a) for details on calculation of standardized logistic regression coefficients and R_L^2 and R_O^2 from SAS and SPSS output.

References

Agresti, A. (2002). *Categorical Data Analysis,* 2nd edn. New York: Wiley.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32: 283–301.

Bijleveld, C. C. J. H. and van der Kamp, L. J. T., with Mooijaart, A., van der Kloot, W. A., van der Leeden, R. and van der Burg, E. (1998). *Longitudinal Data Analysis: Designs, Models, and Methods.* London: Sage.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47: 225–238.

Cronbach, L. J. and Furby, L. (1970). How should we measure change—or should we? *Pychological Bulletin*, 48 (Part 1): 68–80.

Davies, R. B. (1994). From cross-sectional to longitudinal analysis. In A. Dale and R. B. Davies (eds), *Analyzing Social and Political Change: A Casebook of Methods*, pp. 20–40. London: Sage.

Elliott, D. S., Huizinga, D. and Menard, S. (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems.* New York: Springer-Verlag.

Finkel, S. E. (1995). *Causal Analysis with Panel Data.* Thousand Oaks, CA: Sage.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized Estimating Equations.* Boca Raton, FL: Chapman & Hall/CRC.

Heise, D. R. (1975). *Causal Analysis.* New York: Wiley.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression,* 2nd edn. New York: Wiley.

Hsiao, C. (2003). *Analysis of Panel Data,* 2nd edn. Cambridge, UK: Cambridge University Press.

Jöreskog, K. G. and Sörbom, D. (1989). *LISREL 7: A Guide to the Program and Applications,* 2nd edn. Chicago: SPSS, Inc.

Kessler, R. C. and Greenberg, D. G. (1981). *Linear Panel Analysis: Models of Quantitative Change.* New York: Academic Press.

King, G. (1989) *Unifying Political Methodology: The Likelihood theory of Statistical Inference.* Cambridge, UK: Cambridge University Press.

Liker, J. K., Augustyniak, S. and Duncan, G. J. (1985). Panel data and models of change: a comparison of first difference and conventional two-wave models. *Social Science Research*, 12: 80–101.

Maruyama, G. M. (1998). *Basics of Structural Equation Modeling.* Thousand Oaks, CA: Sage.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *American Statistician*, 54: 17–24.

Menard, S. (2002a). *Applied Logistic Regression Analysis.* Thousand Oaks, CA: Sage.

Menard, S. (2002b). *Longitudinal Research.* Thousand Oaks, CA: Sage.

Menard, S. (2004a). Six approaches to calculating standardized logistic regression coefficients. *American Statistician*, 58: 218–223.

Menard, S. (2004b). Path analysis with logistic regression. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Toronto, Canada, (July).

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16: 1–32.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods,* 2nd edn. Thousand Oaks, CA: Sage.

Rogosa, D. (1995). Myths and methods: “Myths about longitudinal research” plus supplemental questions. In J. M. Gottman (ed.), *The Analysis of Change*, pp. 3–66. Mahwah, NJ: Lawrence Erlbaum.

Sayrs, L. W. (1989). *Pooled Time Series Analysis.* Thousand Oaks, CA: Sage.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford, UK: Oxford University Press.

Wooldredge, J. W. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: The MIT Press.

Yamaguchi, K. (1991). *Event History Analysis.* Newbury Park, CA: Sage.

Chapter 31

Latent growth curve models

Michael Stoolmiller

In this chapter, we take up the subject of modeling change over time by building models for growth curves using a structural equation modeling (SEM) approach. We start by defining growth curves and introducing the terminology for describing them. Next, we describe the current available alternatives for fitting growth curve models, their strengths, and limitations. Then we introduce the basic latent growth curve (LGC) model and walk through the fundamental steps of specification, identification, and estimation using an empirical example drawn from developmental psychology. Model identification involves some algebraic manipulations and readers not interested in this level of detail can safely skip these sections. Next, we elaborate the basic LGC model to include preexisting predictors of future growth and repeat the steps of specification, identification, and estimation. Finally, we extend the LGC model again to demonstrate how one can not only include predictors of growth but also growth as a predictor of a distal outcome. The empirical example used throughout illustrates not only a successful application of the methodology but also some of the problems that users are likely to encounter when they begin to fit LGC models.

A growth curve describes how a dependent variable, say y, depends on the passage of time, where time is considered the independent variable. In other words, a growth curve is a function that takes time as input and returns y values. The function has some mathematical form, for example linear, as illustrated in Figure 31.1, and then parameters that control the exact type of the chosen functional form. For example, the parameters for a straight line are intercept and slope, which define a particular straight line and distinguish it from other straight lines with different values of intercept and slope. The intercept of a linear growth curve is defined as the value of the dependent variable, y, when time, the independent variable, is equal to zero, as shown in Figure 31.1 where the line intersects the y axis. The y axis is drawn in this case at the point where time is equal to zero but this is not always true or necessary. The slope of a linear growth curve is defined as the amount of change in the dependent variable, y, for a unit increase in time, the independent variable as shown in Figure 31.1 by the two black arrows.

What is illustrated in Figure 31.1 is a mathematical abstraction, a straight line. In actual practice, it is not usually the case that the dependent variable is continuously observed as implied by the plot in Figure 31.1. More typically, an empirical growth curve consists of a set of repeated measurements of the dependent variable taken at discrete time intervals. Real world data from the social sciences also rarely conform to deterministic mathematical abstractions such as straight lines. More likely,

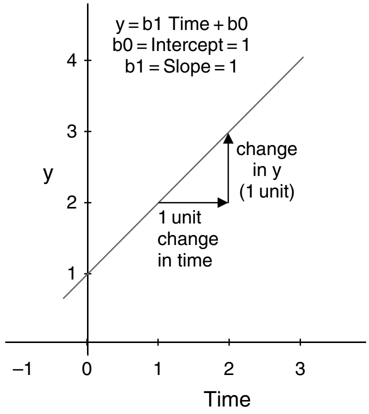


Figure 31.1 Hypothetical growth curve for outcome y

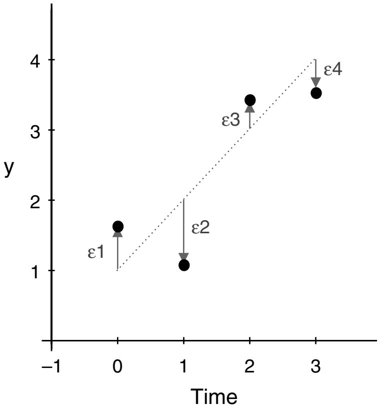


Figure 31.2 Fitted growth curve (dashed line) for outcome y with observed data (black dots). Epsilons (ε) are time-specific influences

a growth curve based on 4 repeated assessments would look like Figure 31.2.

The 4 observed y values are shown by solid black circles. The dashed line is the hypothetical straight line growth curve fit to the observed data. The discrepancies between the observed values and the values implied by the fitted straight line are shown as blue arrows pointing away from the line to the observed data. These discrepancies, denoted as ε1 to ε4 in Figure 31.2, are known as time-specific influences or sometimes also as residual influences. The impact of the time-specific influences is to mask the underlying linear growth curve. These time-specific influences include random measurement error but they could also include predictable influences that operate only at a specific point in time. The observed y scores are equal to the fitted growth curve scores at each time point plus ε at each time point and so they do not follow the nice linear progression as indicated by the fitted line. Note that the y2 score is actually less than the y1 score in Figure 31.2. There is an obvious correspondence of the growth curve in Figure 31.2 to observed variable regression methodology.[1]

In growth curve modeling, it is typically the case that instead of just one growth curve, we have a sample of growth curves, one for each subject. To illustrate how this expands the modeling possibilities, let us assume we have 3 individuals measured at 4 time points and we know their intercept, slope, and observed y values. Their data and fitted growth curves are shown below in Figure 31.3. Suppose individual 1 has an intercept score of 1 and a slope score of 1. This individual's fitted growth curve scores over the 4 times are 1, 2, 3, and 4 and are shown below in the upper left corner of Figure 31.3. Subject 2 has an intercept score of 4 and a slope score of −1 and their fitted values are 4, 3, 2, and 1 and the curve is shown in upper right of Figure 31.3. Subject 3 has an intercept score of 2 and a slope score of 0 and their fitted values are 2, 2, 2, and 2 and the curve is shown in lower left of Figure 31.3. In addition, for each subject, their time-specific values, ε1 to ε4, are shown as arrows pointing away from the

[1] In fact, we could choose to fit the straight line growth curve by regressing the observed y values on the time values using ordinary least squares regression. This approach has the appeal of simplicity and is useful for preliminary graphical analyses as we will demonstrate but it has the disadvantage of being less statistically efficient than fitting growth curves using maximum likelihood techniques.

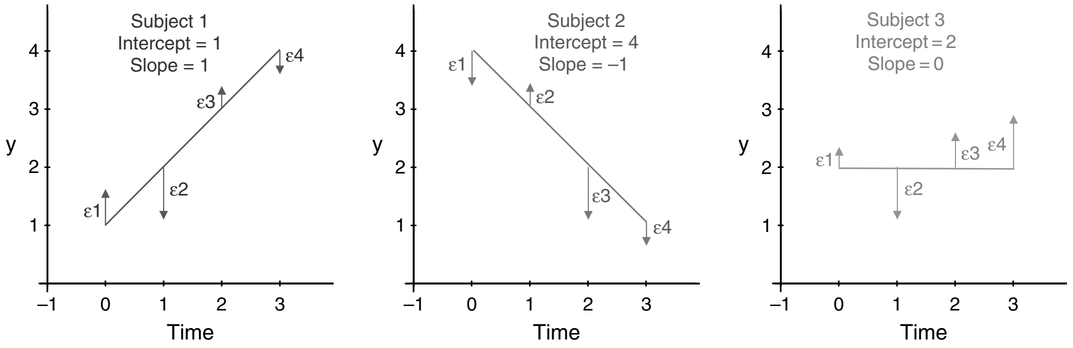


Figure 31.3 Fitted growth curves for 3 individuals (solid lines) for outcome y. Epsilons are time-specific influences

fitted growth curve. The mean of the intercept and slope scores for these 3 subjects would be $(1+4+2)/3 = 2.5$ and $(1-1+0)/2 = 0$, respectively. The variance of the intercept and slope scores for these 3 subjects would be $[(1-2.5)^2+(4-2.5)^2+(2-2.5)^2]/3 = 1.58$ and $[(1-0)^2+(-1-0)^2+(0-0)^2]/3 = 2/3$ respectively. The covariance of the intercept and slope for these 3 subjects is equal to -1.

Now an obvious question to a developmental psychologist is what causes the individual differences in both initial status (intercept scores) and growth rate (slope scores) over time? Why did subject 1 go up? Why did subjects 2 and 3 go down and stay the same, respectively? Did these subjects differ on some important background variable at time zero or earlier that was responsible for the developmental differences in change over time? LGC models can help answer these kinds of questions.

The SEM approach to growth curve models, however, is not the only available alternative. Here we digress briefly to compare the currently available major alternative so that readers can make informed decisions about which approach best meets their needs. The major alternative for fitting growth curve models goes by a variety of terms, including the multilevel model (MLM; e.g., Goldstein, 2003), the random effect model (REM; e.g., Laird and Ware, 1982), the mixed effect model (MEM; e.g., Pinheiro and Bates, 2000) or the hierarchical linear model (HLM; e.g., Raudenbush and Bryk, 2002) depending on author or software package.

There are two important differences that readers may want to consider before embarking on a growth curve analysis. The first major difference between most alternatives and LGC is that LGC is capable of estimating complicated structural models between pre-existing predictors, growth, and future outcomes of growth, including testing for mediation or indirect effects that may flow from pre-existing predictors to growth and then to some distal outcome. In contrast, the alternatives treat growth as the ultimate and only outcome, making it difficult to extend the model to use the growth patterns themselves as predictors or examine indirect effects. The empirical example used in this chapter illustrates this aspect of LGC but for another example, see Stoolmiller, Duncan, Bank and Patterson (1992).

On the other hand, the second major difference is that LGC takes a multivariate approach to repeated measures within subjects, which means that the number of repeated measures and the balance of the repeated assessment design can be a concern, especially in small samples. One strategy to dealing with imbalance in the design (i.e., different time intervals for different individuals) for LGC analyses is to create more repeated assessments but set the data to missing for individuals not observed at that specific interval. As the number of repeated assessments gets large, however, and begins to approach the number of individuals in the data, the estimation procedures can become unstable and ultimately break down. In this situation, multilevel models may be a better choice.[2]

Returning now to the LGC, shown below in Figure 31.4 is a path diagram for a simple linear growth SEM with 4 observed variables that represent repeated assessments of the same measure at 4 time points. Notation follows closely that developed by Muthen and Muthen (2004) for their MPlus package, which was used to estimate all models in this chapter. The model assumes that the time intervals between assessments are the same for all subjects although not all subjects have to have all 4 assessments. Subjects with only 3, 2, or even just 1 assessment can still be included in the model.[3]

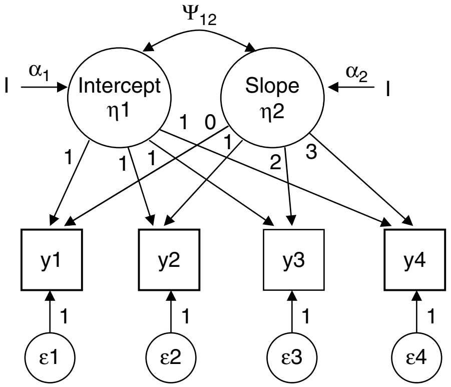


Figure 31.4 Path diagram for linear growth curve model

The model has two latent variables, which are labeled Intercept and Slope. The intercept and slope factors represent the collection of intercepts and slopes for each individual subject's linear growth curve. The intercept and slope have means, α1 and α2, and variances, ψ1 and ψ2, respectively and a covariance, ψ12. The quantities α1 and α2 correspond to the means of the intercept and slope scores, respectively, computed for the example in Figure 31.3 above. The quantities ψ1 and ψ2 correspond to the variances of the intercept and slope scores, respectively, computed for Figure 31.3 and ψ12 corresponds to the covariance. Associated with each observed y measure is a latent, time-specific variable, ε. These correspond to the discrepancies between the fitted growth curve scores and the observed y values as illustrated in Figures 31.2 and 31.3. The factor loadings for the intercept are fixed at 1 and for the slope, the values represent the linear passage of time, 0, 1, 2, and 3. Because of the definition of the intercept of a growth curve as being the value of the outcome when time is zero, the Intercept factor in Figure 31.4 represents individual differences in the outcome, corrected for time-specific influences, at y1, the first time data was collected. If a different set of factor loadings had been employed, the intercept factor would

[2] Mplus, the program used for all analyses in this chapter, is an exception to this rule and can handle highly imbalanced data in the same manner as the alternative programs.

[3] With longitudinal data, it is often the case that a substantial portion of the sample has only partial data and a model that required every subject to have all 4 assessments would not be very useful. Most SEM programs have options for using cases with partial data and this option should be the default choice. Using cases with partial data reduces potential biases, maximizes statistical power and is almost always superior to older ad hoc approaches to dealing with missing data, including using only subjects with complete data or subjects with some minimum number of assessments (e.g., 2 out of 4) (Schaffer and Graham, 2002). In fact, so long as the fact that the data are missing does not depend on the value of the missing data, estimation can proceed using cases with partial data without introducing bias relative to the target population from which the sample was drawn. This condition is usually referred to as ignorable missingness. Keep in mind, however, that even if ignorable missingness does not hold, using only subjects with complete data or some minimum number of assessments will usually result in even more bias.

represent true individual differences at a different time. Say, for example, the factor loadings of −2, −1, 0 and 1 were used. Then the intercept factor would represent true individual differences at y3, the third data collection point. Or if the factor loadings of −3, −1, 1 and 3 were used, the intercept factor would represent individual differences midway between time points 2 and 3, a time point for which no data was actually collected. This particular point is actually the mean of the 4 fitted values over time, which makes it an interesting choice for an intercept factor. The fact that it is midway between y2 and y3 amounts to an interpolation but this is reasonable given our choice of a linear growth model. Another good choice for factor loadings could be −3, −2, −1 and 0 which would shift the intercept to representing true individual differences at the last point of observation. The choice of scaling for the slope, and hence the interpretation of the intercept, can be determined by the substantive goals of the research. Choosing factor loadings such that the intercept is beyond the range of the observed data is less desirable because interpreting the results becomes more problematic. For example, suppose the factor loadings were 1, 2, 3, and 4. The intercept factor in this case represents true individual differences at time 0 which is 1 year prior to when any data were collected. Extrapolating linear models beyond the range of the data is risky and is not usually advisable.

We can make all of this more precise by writing down the structural equations for the model in Figure 31.4 for the ith individual. There is one equation for each dependent variable, y1 to y4,

$$\begin{aligned} y1_i &= \eta1_i + 0\eta2_i + \varepsilon1_i \\ y2_i &= \eta1_i + 1\eta2_i + \varepsilon2_i \\ y3_i &= \eta1_i + 2\eta2_i + \varepsilon3_i \\ y4_i &= \eta1_i + 3\eta2_i + \varepsilon4_i. \end{aligned} \tag{1}$$

Note how the intercept, η1, makes a constant contribution to the y scores but the slope, η2, makes a contribution that is scaled by the fixed factor loadings according to the linear passage of time. Note too that η1 just represents individual differences at y1, corrected for time-specific influences. In the most elementary and standard LGC, and the only one discussed in this chapter, the latent variables are all assumed to be jointly multivariate normally distributed, which implies that the y's are all multivariate normally distributed. LGC methods, however, have been extended to accommodate a variety of different types of variables (e.g., dichotomous, ordered categorical, count, and censored variables; see Muthen and Muthen, 2004).

Now that we have specified the model, the next task is to show that the model parameters are identified.[4] The means[5] of the y variables are

$$\begin{aligned} E(y1) &= E(\eta1 + 0\eta2 + \varepsilon1) \\ &= E(\eta1) \\ &= \alpha1 \end{aligned} \tag{2}$$

$$\begin{aligned} E(y2) &= E(\eta1 + 1\eta2 + \varepsilon2) \\ &= E(\eta1) + E(\eta2) \\ &= \alpha1 + \alpha2. \end{aligned} \tag{3}$$

From equation (2) it is easy to see that α1 is identified and in fact is just equal to the mean of y1. Given that α1 is identified, then α2 is easily identified from equation (3) and in fact

[4] Identification is the process of demonstrating that the available data is sufficient to uniquely determine each parameter in the model. For LGC, the available data consists of means, variances, and covariances of the observed variables. Thus, identification is showing that the model parameters can be uniquely determined by the means, variances, and covariances of the observed variables.

[5] Upper case E indicates the mathematical operation of computing expectation or the mean of a random variable. For a review of the rules of expectation, see Kirk (1982).

is just the mean of y2 minus the mean of y1. The mean structure does not involve any other model parameters so we must turn to the covariance structure to identify the variances and covariances of the latent variables.[6] The variance of y1 is

$$\begin{aligned}\operatorname{var}(y1) &= \operatorname{var}(\eta1+\varepsilon1)\\ &= \operatorname{var}(\eta1)+\operatorname{var}(\varepsilon1)+2\operatorname{cov}(\eta1,\varepsilon1) \qquad (4)\\ &= \operatorname{var}(\eta1)+\operatorname{var}(\varepsilon1).\end{aligned}$$

The covariance of y1 with y2 is

$$\begin{aligned}\operatorname{cov}(y1,y2) &= \operatorname{cov}(\eta1+\varepsilon1,\eta1+\eta2+\varepsilon2)\\ &= \operatorname{cov}(\eta1,\eta1)+\operatorname{cov}(\eta1,\eta2)\\ &\quad +\operatorname{cov}(\eta1,\varepsilon2)+\operatorname{cov}(\varepsilon1,\eta1) \qquad (5)\\ &\quad +\operatorname{cov}(\varepsilon1,\eta2)+\operatorname{cov}(\varepsilon1,\varepsilon2)\\ &= \operatorname{var}(\eta1)+\operatorname{cov}(\eta1,\eta2).\end{aligned}$$

The covariance of y1 with y3 is

$$\begin{aligned}\operatorname{cov}(y1,y3) &= \operatorname{cov}(\eta1+\varepsilon1,\eta1+2\eta2+\varepsilon3)\\ &= \operatorname{cov}(\eta1,\eta1)+2\operatorname{cov}(\eta1,\eta2)\\ &\quad +\operatorname{cov}(\eta1,\varepsilon3)+\operatorname{cov}(\varepsilon1,\eta1) \qquad (6)\\ &\quad +2\operatorname{cov}(\varepsilon1,\eta2)+\operatorname{cov}(\varepsilon1,\varepsilon3)\\ &= \operatorname{var}(\eta1)+2\operatorname{cov}(\eta1,\eta2).\end{aligned}$$

If we subtract equation (6) from equation (5) we get

$$\begin{aligned}&\operatorname{cov}(y1,y3)-\operatorname{cov}(y1,y2)\\ &\quad = \operatorname{var}(\eta1)+2\operatorname{cov}(\eta1,\eta2)-\operatorname{var}(\eta1)\\ &\qquad +\operatorname{cov}(\eta1,\eta2) \qquad (7)\\ &\quad = \operatorname{cov}(\eta1,\eta2),\end{aligned}$$

which identifies the covariance of $\eta1$ and $\eta2$, $\psi12$. Back substituting the value of the covariance of $\eta1$ and $\eta2$ into equation (5) we get

$$\begin{aligned}\operatorname{cov}(y1,y2) &= \operatorname{var}(\eta1)+\operatorname{cov}(\eta1,\eta2)\\ &= \operatorname{var}(\eta1)+\operatorname{cov}(y1,y3) \qquad (8)\\ &\quad -\operatorname{cov}(y1,y2)\\ 2\operatorname{cov}(y1,y2)-\operatorname{cov}(y1,y3) &= \operatorname{var}(\eta1),\end{aligned}$$

which identifies the variance of $\eta1$. The variance of $\varepsilon1$ is then easily identified by equation (4). The covariance of y2 with y3 is

$$\begin{aligned}\operatorname{cov}(y2,y3) &= \operatorname{cov}(\eta1+\eta2+\varepsilon2,\eta1+2\eta2+\varepsilon3)\\ &= \operatorname{cov}(\eta1,\eta1)+2\operatorname{cov}(\eta1,\eta2)+\operatorname{cov}(\eta1,\varepsilon3)\\ &\quad +\operatorname{cov}(\eta2,\eta1)+2\operatorname{cov}(\eta2,\eta2)+\operatorname{cov}(\eta2,\varepsilon3)\\ &\quad +\operatorname{cov}(\varepsilon2,\eta1)+2\operatorname{cov}(\varepsilon2,\eta2)+\operatorname{cov}(\varepsilon2,\varepsilon3)\\ &= \operatorname{var}(\eta1)+3\operatorname{cov}(\eta1,\eta2)+2\operatorname{var}(\eta2). \qquad (9)\end{aligned}$$

With the variance of $\eta1$ and covariance of $\eta1$ and $\eta2$ already identified, equation (9) identifies the variance of $\eta2$. The variance of y2 is

$$\begin{aligned}\operatorname{var}(y2) &= \operatorname{var}(\eta1+\eta2+\varepsilon2)\\ &= \operatorname{var}(\eta1)+\operatorname{var}(\eta2)+\operatorname{var}(\varepsilon2) \qquad (10)\\ &\quad +2\operatorname{cov}(\eta1,\eta2).\end{aligned}$$

The variance of $\varepsilon2$ is the only unidentified parameter in equation (10) so the variance of $\varepsilon2$ is identified. Similarly, the variances of y3 and y4 identify the variances of $\varepsilon3$ and $\varepsilon4$. Thus, the model is completely identified. It is in fact, overidentified because we have 14 total degrees of freedom from the 4 means, 4 variances and 6 covariances for the y's and 9 estimated parameters leaving 5 degrees of freedom to test the fit of the model. The entire mean and covariance structure is shown below in Table 31.1.

To demonstrate growth curve modeling, we will model growth in Deviant Peer Affiliation (DPA) in the Oregon Youth Study (OYS) from

[6] The var and cov operators indicate mathematical operations of computing variance or covariance for random variables. For a review of the rules of variance or covariance, see Kirk (1982).

Table 31.1 Mean and covariance structure for 4 points in time linear growth model

	$y1$	$y2$	$y3$	$y4$
y1	$\Psi_{11} + \Theta_{11}$			
y2	$\Psi_{11} + \Psi_{12}$	$\Psi_{11} + \Psi_{22} + 2\Psi_{12} + \Theta_{22}$		
y3	$\Psi_{11} + 2\Psi_{12}$	$\Psi_{11} + 2\Psi_{22} + 3\Psi_{12}$	$\Psi_{11} + 4\Psi_{22} + 4\Psi_{12} + \Theta_{33}$	
y4	$\Psi_{11} + 3\Psi_{12}$	$\Psi_{11} + 3\Psi_{22} + 4\Psi_{12}$	$\Psi_{11} + 6\Psi_{22} + 5\Psi_{12}$	$\Psi_{11} + 9\Psi_{22} + 6\Psi_{12} + \Theta_{44}$
Means	α_1	$\alpha_1 + \alpha_2$	$\alpha_1 + 2\alpha_2$	$\alpha_1 + 3\alpha_2$

Notes: $\Psi_{11} = \text{var}(\eta_1), \Psi_{22} = \text{var}(\eta_2), \Psi_{12} = \text{cov}(\eta_1, \eta_2), \alpha_1 = \text{E}(\eta_1), \alpha_2 = \text{E}(\eta_2), \Theta_{11} = \text{var}(\varepsilon_1), \Theta_{22} = \text{var}(\varepsilon_2), \Theta_{33} = \text{var}(\varepsilon_3), \Theta_{44} = \text{var}(\varepsilon_4)$

Grade 4 to 8. The OYS started in 1984 with 10-year-old boys and their families and is a predominantly white, working-class sample from high-risk neighborhoods in a mid-size city in Oregon. Details on the sample and measures can be found in Stoolmiller (1994). DPA has been implicated as a key proximal predictor of delinquent behavior in adolescence from a number of theoretical perspectives. Thus it is interesting to determine how DPA develops prior to adolescence and identify childhood predictors of growth in DPA. Such predictors may in turn be important targets for interventions aimed at reducing adolescent delinquent behavior. One such set of predictors is father variables and in particular father socioeconomic status (SES). Thus, our example will be limited to the subset of 141 boys in the OYS who had a father figure available to participate in the project at grade 4. We measured DPA in the OYS by child self-report, teacher report, and parent report taken when the boys were in grades 4, 6, 7, and 8. The two-year gap between the first two assessments departs from a simple linear progression but this can easily be adjusted in the model by using similarly scaled factor loadings such as 0, 2, 3, and 4.

The parent and teacher measures were individual items measured on a 3 point scale of 0, 1 and 2. The child report items, however, were measured on 5 point scales of 1, 2, 3, 4, and 5. Arbitrary scaling differences such as these can result in the measure with the largest scale dominating the final construct score. To prevent this, a simple recoding scheme was employed that preserved the raw information necessary to model change over time. The child items were recoded to have zero as a bottom scale anchor by subtracting 1 and then were rescaled to have an upper anchor of 2 by multiplying by $^1/_2$. With this adjustment, the final DPA construct score was the average of the parent, teacher, and child measures, square root transformed to reduce skewness. For modeling purposes, the DPA scores were also multiplied by 100 to provide a more convenient scaling. A common practice with this type of data is to standardize all items at all points in time to a mean of zero and variance of 1 before computing construct scores. Clearly, this practice will not work for growth curve analysis because all information about mean level change and most of the information about variance shifts is discarded.

The observed DPA trajectories are displayed in Figure 31.5, sorted by average level and then linear trend within average level. The sorting helps to group similar trajectories in the same subpanel and leads to a better display that can be helpful in determining the basic functional form that will be necessary to model the trajectories. In this case, it appears as though straight lines would do a reasonably good job of representing the individual trajectories. It is also apparent that a number of subjects increased

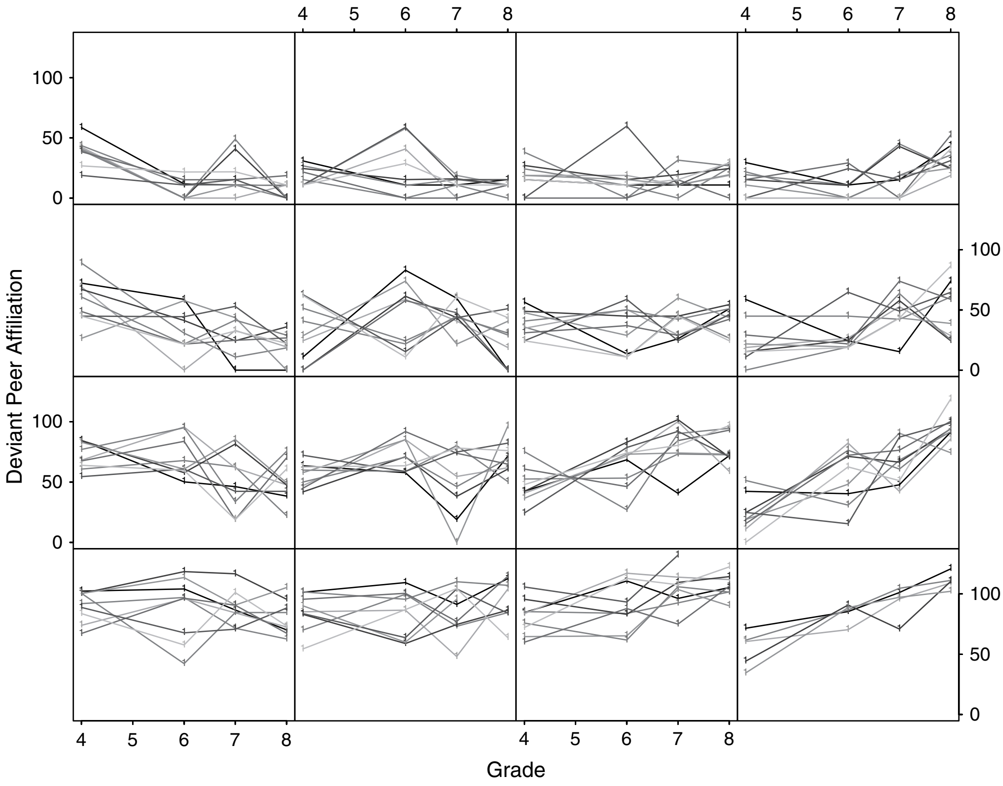


Figure 31.5 Observed individual growth curves for Deviant Peer Affiliation (DPA) from grade 4 to 8

substantially over time (subplots in the bottom two rows, far right column) and some subjects decreased over time (subplots in the middle two rows, far left column). Thus, there would appear to be individual differences in change to be explained.

The scatterplot matrix of the 4 repeated assessments is shown in Figure 31.6. The matrix has normal quantile plots on the main diagonal to help determine if the variables follow a Gaussian distribution. As can be seen, there is some evidence of a floor effect, a number of cases all clumped at the bottom of the scale, at each time point, but the number is not excessive. In addition, skewness (denoted sk in the top margin of the normal quantile plot) is minimal and kurtosis (denoted k in the top margin of the normal quantile plot) is modest but consistently negative indicating lighter tails than the Gaussian distribution. No consistent nonlinearities appear in any of the scatterplots as indicated by the scatterplot smoother (solid line), which tracks the fitted linear regression line (dashed line) quite closely. Thus, there is no indication in the data that modeling should not proceed.

The means and standard deviations, which are printed in the top margin of the normal quantile plots (m denotes mean, sd denotes standard deviation), increase steadily from grade 4 to 8. Individual differences in linear trends imply changing variances over time so the fact that the standard deviations change (i.e., increase) suggests that there are individual differences in change to explain, consistent with the plot of the trajectories in Figure 31.5.

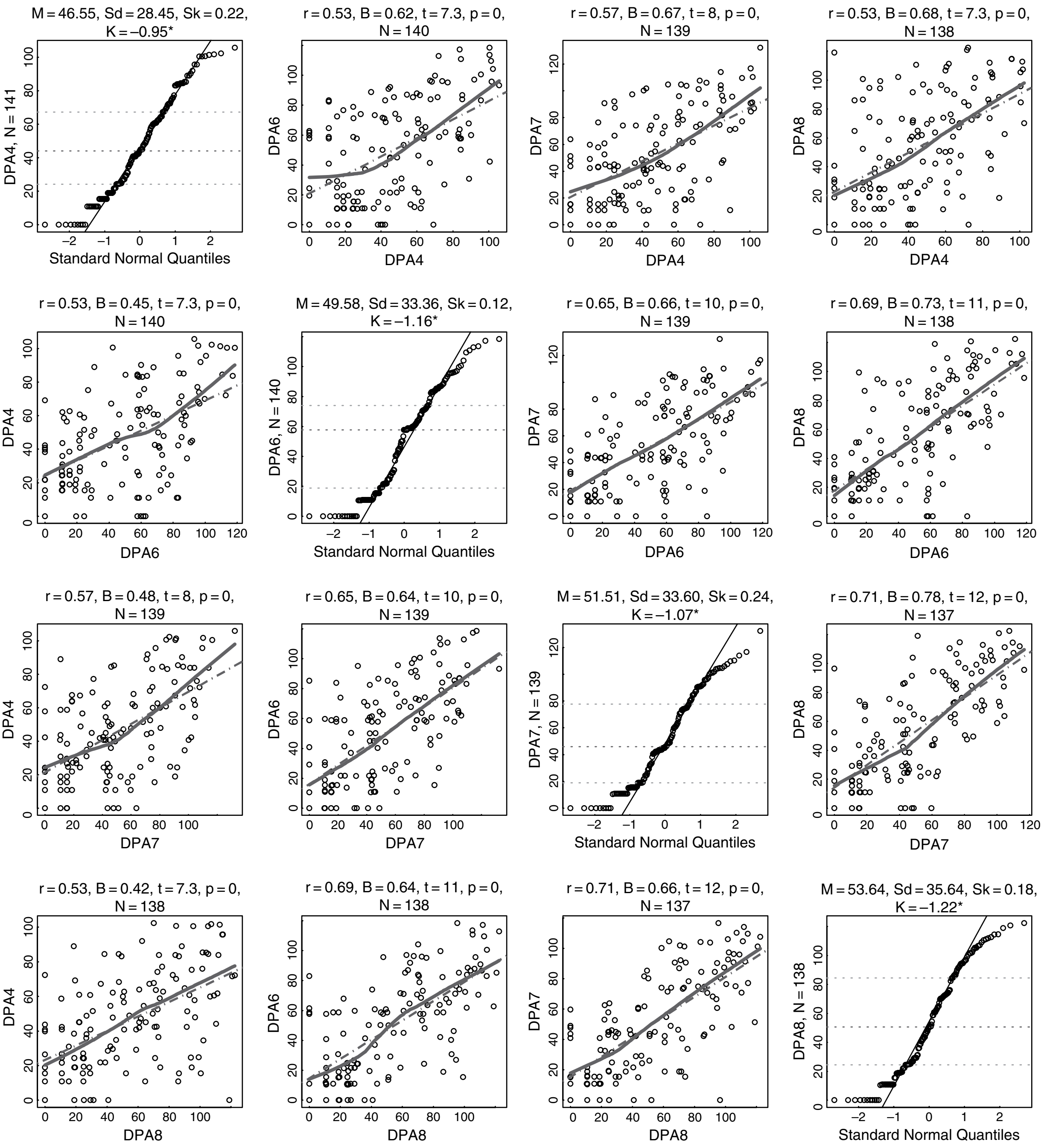


Figure 31.6 Scatterplot matrix for repeated assessments of Deviant Peer Affiliation, grade 4 to 8 (M = mean, sd = standard deviation, sk = skewness, k = kurtosis, r = correlation, b = regression weight, t = t test, p = p level of t test, N = sample size) with normal quantile plots on the main diagonal

The result of fitting the following linear growth curve model:

$$\begin{aligned} y1_i &= \eta 1_i + \varepsilon 1_i \\ y2_i &= \eta 1_i + 2\eta 2_i + \varepsilon 2_i \\ y3_i &= \eta 1_i + 3\eta 2_i + \varepsilon 3_i \\ y4_i &= \eta 1_i + 4\eta 2_i + \varepsilon 4_i \end{aligned} \tag{11}$$

is shown in Table 31.2 (labeled model 1) along with two additional growth models that are simplifications of model 1. Note that the factor loadings were set to match the spacing of the assessments. In addition, the variances of the time-specific influences were constrained to be equal, a common simplifying assumption.

The model 1 chi-square is 3.51 with 8 degrees of freedom (p = .90) and the nonsignificant p value indicates that the model implied covariance matrix and mean vector adequately reproduces the observed counterparts. The chi-square is a "badness of fit" test statistic that for a fixed sample size gets bigger as the model and data become increasingly discrepant. For a fixed degree of model fit, the chi-square also gets bigger as the sample size increases much the same as a t test for a difference in means increases with sample size. Thus even small discrepancies between the model and the data become "significant" given a large enough sample. A number of fit indices that are not functions of the sample size are also shown in Table 31.2, the comparative fit index (CFI), the Tucker-Lewis fit index (TLI) and the root mean square error of approximation (RMSEA). The CFI and TLI will both be close to 1 for models that fit well and both are essentially 1 for model 1. The RMSEA will be .05 or less for models that fit well and it is essentially zero for model 1. By all indications, model 1 provides an excellent fit.

The parameter estimates are also shown in Table 31.2. The variance of the intercept is strongly significant (critical ratio greater than 1.96 or less than −1.96 implies $p < .05$ by a two-tailed test) but the variance of the slope

Table 31.2 Parameter estimates (Est.), standard errors (SE) and critical ratios (CR) for DPA growth models

	Model 1			*Model* 2			*Model* 3		
	Est.	*SE*	*CR*	*Est.*	*SE*	*CR*	*Est.*	*SE*	*CR*
Means									
DPA intercept	46.26	2.30	20.11	46.26	2.42	19.14	46.28	2.67	17.34
DPA slope	1.93	0.64	3.02	1.92	0.67	2.86	1.90	0.58	3.28
cov(DPA intercept, DPA slope)	35.87	19.89	1.80						
Variances									
DPA intercept	440.05	92.51	4.76	531.77	85.08	6.25	662.44	91.71	7.22
DPA slope	14.99	7.69	1.95	22.80	6.74	3.39			
Residual variances									
DPA4-DPA8	367.91	31.19	11.80	351.34	27.31	12.86	411.73	28.51	14.44
Chi-square	3.51			6.36			24.09		
DF	8.00			9.00			10.00		
P value	0.90			0.70			0.01		
CFI	1.00			1.00			0.95		
TLI	1.01			1.01			0.97		
RMSEA 90% CI (lower limit, estimate, upper limit)	0.00	0.00	0.04	0.00	0.00	0.07	0.05	0.10	0.15

is only marginally significant. This is not particularly surprising for the intercept because it represents true individual differences at time zero which is the first assessment point. For the slope, however, the marginally significant variance implies that strong evidence is lacking for individual differences in change beyond what would be expected by sampling variation alone (i.e., that are potentially explainable).[7] The covariance of the slope and intercept is not significant although it standardizes to a fairly substantial correlation of .44, a point we will return to shortly. This means that there is no association between where a boy starts in grade 4 and how he will change over time. For any given starting point, he is just as likely to go up as to go down.

The mean of the intercept factor is 46.3 and significant but again this is not particularly interesting because it is essentially a test that the mean of the observed DPA measure at the first assessment is not zero. Unless normative data were available for these DPA measures, this is not a particularly useful fact. The mean of the slope factor, however, is 1.93 and significantly different from zero. In other words, although some individuals are going up and some are going down, overall for the entire sample, there is an increase in DPA from grade 4 to 8 of about $1.93^*4 = 7.72$. This suggests that for the population, as boys make the transition from elementary to middle school, they tend to associate with peers who are more deviant. The mean shift of 7.72 from grade 4 to grade 8 is about .37 of the standard deviation of the intercept, suggesting a medium effect size for population level growth.

The residual variance is 367.91 at each point in time which implies that the intercept and slope factors account for 55, 64, 68, and 72% of the variance of the observed measures at the first through fourth points of assessment respectively. If the variance of the observed measures is increasing due to the individual differences in change over time, and if the residual variances are constrained to be equal across time, then the R^2 for the individual observed variables must go up. Intuitively, this means that the growth process is creating increasing amounts of true score variance relative to error variance, so the reliability of the measures, the R^2, must go up.

Although the estimates all seem reasonable, are within the admissible parameter space,[8] and do not show troublesome levels of

[7] If the test for the variance of the slope was not significant, and it is good to keep in mind that tests of variance parameters being zero in SEM tend to be biased against the nonzero alternative (see Pinheiro and Bates, 2000, pp. 84–87), it would suggest that the apparent change visible in Figure 31.4 is due to chance variation and may not be explainable. It is still possible that small amounts of slope variation that are apparently nonsignificant by a biased statistical test are nonetheless due to some systematic cause. If it was hypothesized a priori and verified that some predictor significantly predicted the Slope factor, this would be evidence against the hypothesis of chance variation.

[8] The admissible parameter space refers to the set of parameters that make statistical sense, that is, non-negative variances and correlations between −1 and 1. Most SEM programs will return values outside of this range if they happen to maximize the fit of the model to the data and these are usually referred to as improper solutions. An improper solution may indicate a serious problem such as an unidentified or badly misspecified model so they should be carefully investigated. On the other hand, an improper solution may also arise because of sampling variation and a population parameter that is truly close to or on the boundary of the admissible parameter space. In this case, the parameter should not be significantly different from the boundary value. Keep in mind, however, that standard tests involving boundary values for variance parameters can be biased against the non-boundary alternative, especially in modest samples. See Pinheiro and Bates (2000, pp. 84–87) for more discussion.

confounding (the largest correlation among the parameter estimates, which are not shown, was −.50), the fact that the slope variance is only marginally significant and the intercept–slope correlation despite being .44 is not significant merits a closer look. Model 2 in Table 31.2 shows the result of dropping the intercept–slope covariance from the model. A nested chi-square test comparing model 2 to model 1 is 2.85 with 1 degree of freedom, $p = .09$, indicating that dropping the intercept–slope covariance does not cause a significant degradation in fit consistent with lack of significance of the same parameter in model 1. Interestingly, the slope variance in model 2 is now substantially larger (about 50% increase), the slope variance standard error drops slightly and the critical ratio, 3.39, is now strongly significant, $p < .001$. Model 3 shows results when the slope variance is dropped from the model and the nested chi-square test for comparing model 2 to model 3 is 17.73 with 1 degree of freedom, highly significant and consistent with the strong significance of the same parameter in model 2. From this series of model fits, we see that model 1 is misleading because the data is somewhat consistent with a positive intercept–slope covariance (e.g., the increasing covariances between grade 4 DPA and grades 6, 7, and 8 DPA can only come from a positive intercept–slope covariance, see Table 31.1) but the remaining covariances are not big enough to support the significance of the intercept–slope covariance and the slope variance simultaneously. Once the intercept–slope covariance is eliminated, however, the slope variance is strongly significant. Fortunately, dropping the slope variance and keeping the intercept–slope covariance does not make sense because variation is a prerequisite for covariation, so the choice among models 1, 2, and 3 is straightforward and we will build on model 2 for the rest of the models in this chapter.

Building a growth model is the first step toward answering the question posed earlier about why some subjects go up, others go down, and still others stay the same across time. Adding covariates to the model helps answer the question. When the covariates are measured at or before the first time period, a significant effect on the slope is powerful evidence that they are implicated in the growth process. If the covariates are measured somewhere in the middle or towards the end of the developmental period under study, it becomes less clear whether they are a determinant or consequence of growth. For DPA, family, child and parent attributes are obvious choices as determinants of future growth. For our purposes here, we will consider grade 4 covariates of the boy's academic skill and the father's SES as potential predictors. Before we actually fit the model, we first examine the structural equations and check for identification problems.

Suppose the basic growth curve model, illustrated in Figure 31.4, is augmented to

$$\begin{aligned}
y1_i &= \eta1_i + 0\eta2_i + \varepsilon1_i\\
y2_i &= \eta1_i + 1\eta2_i + \varepsilon2_i\\
y3_i &= \eta1_i + 2\eta2_i + \varepsilon3_i\\
y4_i &= \eta1_i + 3\eta2_i + \varepsilon4_i\\
\eta1_i &= \Gamma11x1_i + \zeta1_i + \alpha1\\
\eta2_i &= \Gamma21x1_i + \zeta2_i + \alpha2
\end{aligned} \tag{12}$$

The model is shown below in Figure 31.7. Notice now that the intercept and slope factors have become dependent variables, predicted by x1, and both now have residual variances and intercepts. There is also a covariance between the growth factor residuals to reflect other unmeasured causes that might cause the intercept and slope factors to be correlated. We have previously shown that the means and variances of the intercept and slope factors were identified. If we can show that any new parameters that go into making up the means and variances of the intercept and slope factors are identified (i.e., Γ11 and Γ21), we can use our previous

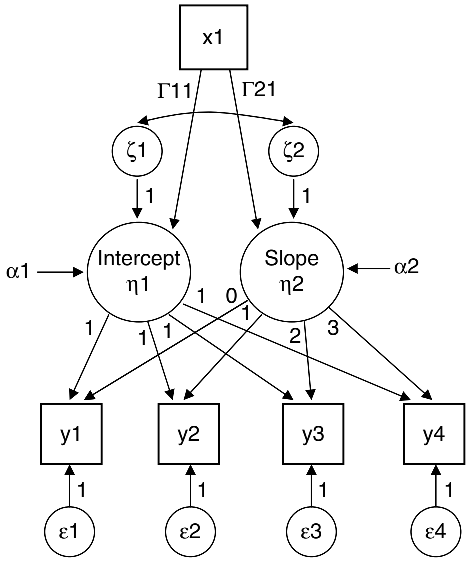


Figure 31.7 Path diagram for linear growth curve model with pre-existing predictor of intercept and slope

results to argue that the rest of the model is identified once Γ11 and Γ21 are identified.

The mean and variance of x1 are of course just taken as their sample values since x1 is an independent variable. The covariance of x1 with y1 is

$$
\begin{aligned}
\mathrm{cov}(x1, y1) &= \mathrm{cov}(x1, \eta 1 + 0\eta 2 + \varepsilon 1)\\
&= \mathrm{cov}(x1, \Gamma 11 x1 + \zeta 1 + \alpha 1 + \varepsilon 1)\\
&= \mathrm{cov}(x1, \Gamma 11 x1 + \zeta 1 + \varepsilon 1)\\
&= \mathrm{cov}(x1, \Gamma 11 x1) + \mathrm{cov}(x1, \zeta 1)\\
&\quad + \mathrm{cov}(x1, \varepsilon 1)\\
&= \Gamma 11 \mathrm{var}(x1) \qquad (13)
\end{aligned}
$$

In step 2 of equation (13), the definition of η1 was substituted in and η2 was dropped because it is multiplied by zero. In step 3, α1 was dropped since it is a constant and therefore cannot contribute to covariation. What equation (13) shows is that Γ11 is identified because var(x1) is a given. The covariance of x1 with y2 is

$$
\begin{aligned}
\mathrm{cov}(x1, y2) &= \mathrm{cov}(x1, \eta 1 + \eta 2 + \varepsilon 2)\\
&= \mathrm{cov}(x1, \Gamma 11 x1 + \zeta 1 + \alpha 1 + \Gamma 21 x1 + \zeta 2\\
&\quad + \alpha 2 + \varepsilon 2)\\
&= \mathrm{cov}(x1, \Gamma 11 x1 + \zeta 1 + \Gamma 21 x1 + \zeta 2 + \varepsilon 2)\\
&= \mathrm{cov}(x1, \Gamma 11 x1) + \mathrm{cov}(x1, \zeta 1)\\
&\quad + \mathrm{cov}(x1, \Gamma 21 x1) + \mathrm{cov}(x1, \zeta 2) + \mathrm{cov}(x1, \varepsilon 2)\\
&= \Gamma 11 \mathrm{var}(x1) + 0 + \Gamma 21 \mathrm{var}(x1) + 0 + 0\\
&= (\Gamma 11 + \Gamma 21)\mathrm{var}(x1) \qquad (14)
\end{aligned}
$$

As in the previous equation, we substituted the definitions of η1 and η2 in and dropped the constants. Now, since var(x1) is a given and Γ11 is already identified, Γ21 is identified. Thus the model is completely identified and in fact, overidentified because we have $5(5+3)/2 = 20$ total degrees of freedom minus 13 estimated parameters, leaving 7 degrees of freedom to test the model.

Returning now to our empirical example, Figure 31.8 shows grade 4 academic skill and father SES against each subsequent DPA measure. This is a powerful graphical technique for identifying early predictors of future change. This can be seen by computing the regression weight for the early predictor against each repeated assessment of the outcome assuming the model in Figure 31.7 holds. These regression weights are the covariance of the outcome, each successive DPA measure, y1 through y4, and the predictor, x1, divided by the variance of the predictor, x1. The regression weight for y1 is

$$
B_{x1,y1} = \frac{\mathrm{cov}(x1, y1)}{\mathrm{var}(x1)} = \frac{\Gamma 11 \mathrm{var}(x1)}{\mathrm{var}(x1)} = \Gamma 11 \qquad (15)
$$

The regression weight for y2 is

$$
\begin{aligned}
B_{x1,y2} &= \frac{\mathrm{cov}(x1, y2)}{\mathrm{var}(x1)} = \frac{(\Gamma 11 + \Gamma 21)\mathrm{var}(x1)}{\mathrm{var}(x1)} \qquad (16)\\
&= \Gamma 11 + \Gamma 21
\end{aligned}
$$

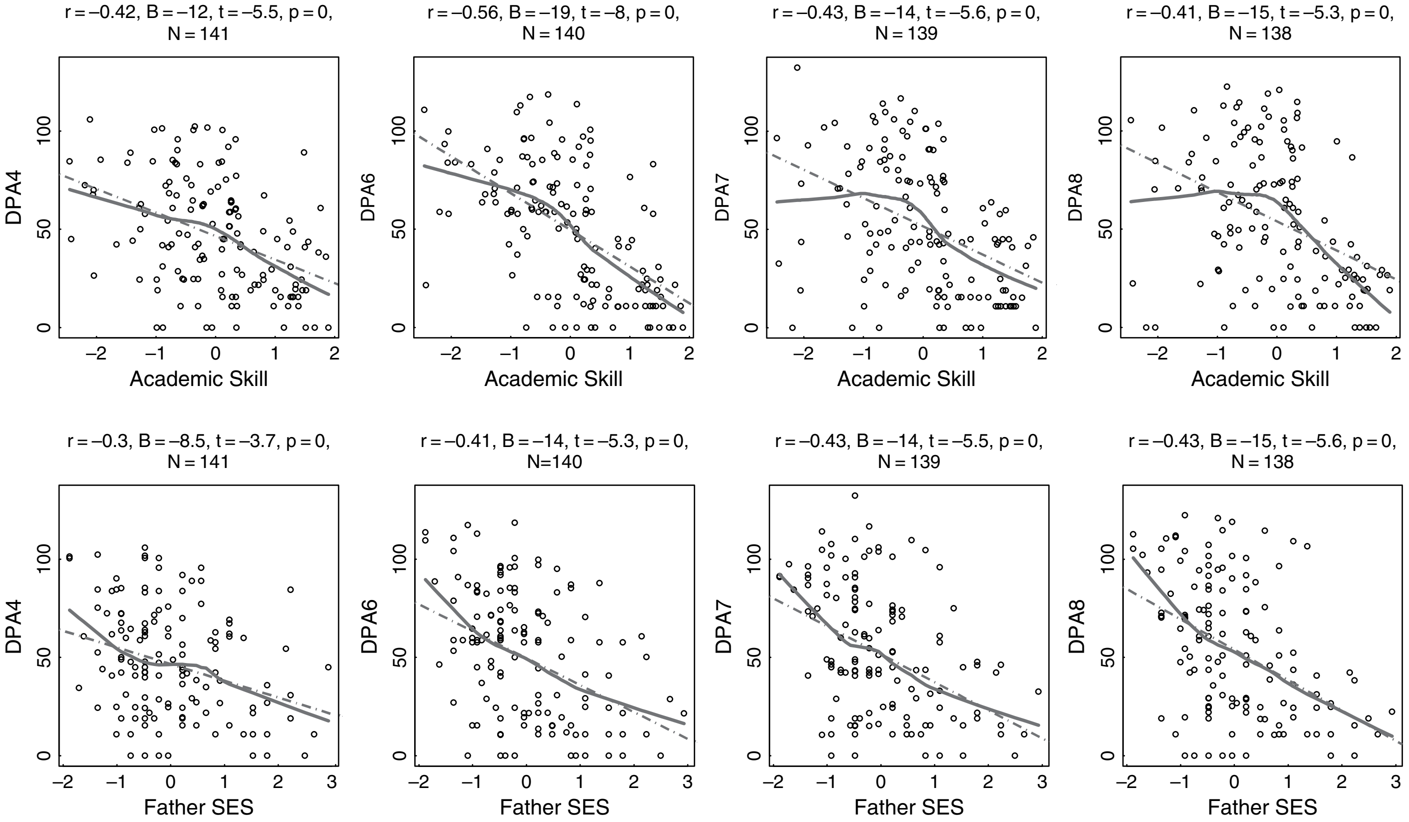


Figure 31.8 Scatterplots of grade 4 academic skill and father SES versus repeated assessments of DPA, grade 4 to 8

To get the regression weight for y3, we first compute the covariance of x1 and y3.

$$\begin{aligned}\mathrm{cov}(x1,y3) &= \mathrm{cov}(x1,\eta1+2\eta2+\varepsilon3)\\ &= \mathrm{cov}(x1,\Gamma11x1+\zeta1+\alpha1+2\Gamma21x1\\ &\quad +2\zeta2+2\alpha2+\varepsilon3)\\ &= \mathrm{cov}(x1,\Gamma11x1+\zeta1+2\Gamma21x1+2\zeta2+\varepsilon3)\\ &= \mathrm{cov}(x1,\Gamma11x1)+\mathrm{cov}(x1,\zeta1)\\ &\quad +\mathrm{cov}(x1,2\Gamma21x1)+\mathrm{cov}(x1,2\zeta2)\\ &\quad +\mathrm{cov}(x1,\varepsilon3)\\ &= \Gamma11\mathrm{var}(x1)+0+2\Gamma21\mathrm{var}(x1)+0+0\\ &= (\Gamma11+2\Gamma21)\mathrm{var}(x1) \qquad (17)\end{aligned}$$

The regression weight is

$$\begin{aligned}B_{x1,y3} = \frac{\mathrm{cov}(x1,y3)}{\mathrm{var}(x1)} &= \frac{(\Gamma11+2\Gamma21)\mathrm{var}(x1)}{\mathrm{var}(x1)}\\ &= \Gamma11+2\Gamma21 \qquad (18)\end{aligned}$$

To get the regression weight for y4, we first compute the covariance of x1 and y4.

$$\begin{aligned}\mathrm{cov}(x1,y4) &= \mathrm{cov}(x1,\eta1+3\eta2+\varepsilon4)\\ &= \mathrm{cov}(x1,\Gamma11x1+\zeta1+\alpha1+3\Gamma21x1+3\zeta2\\ &\quad +3\alpha2+\varepsilon4)\\ &= \mathrm{cov}(x1,\Gamma11x1+\zeta1+3\Gamma21x1+3\zeta2+\varepsilon4)\\ &= \mathrm{cov}(x1,\Gamma11x1)+\mathrm{cov}(x1,\zeta1)\\ &\quad +\mathrm{cov}(x1,3\Gamma21x1)+\mathrm{cov}(x1,3\zeta2)\\ &\quad +\mathrm{cov}(x1,\varepsilon4)\\ &= \Gamma11\mathrm{var}(x1)+0+3\Gamma21\mathrm{var}(x1)+0+0\\ &= (\Gamma11+3\Gamma21)\mathrm{var}(x1) \qquad (19)\end{aligned}$$

The regression weight is

$$\begin{aligned}B_{x1,y4} = \frac{\mathrm{cov}(x1,y4)}{\mathrm{var}(x1)} &= \frac{(\Gamma11+3\Gamma21)\mathrm{var}(x1)}{\mathrm{var}(x1)}\\ &= \Gamma11+3\Gamma21 \qquad (20)\end{aligned}$$

It is easy to see that the bivariate regression weights start at Γ11 and then grow linearly over time by the amount Γ21 for each unit of time if the model in Figure 31.7 holds. Of course, we would expect sampling variation to mask, somewhat, the nice linear progression but overall the regression weights should tend on average to change linearly. If the regression weights do not systematically increase it suggests that Γ21 is zero or, in other words, the variable does not predict linear growth. If the weights seem to change in some systematic pattern but not linearly, it suggests that the predictor predicts nonlinear growth. If the regression at each point in time is nonlinear but changes systematically over time, it suggests the predictor has a nonlinear relation to linear or possibly nonlinear growth in the outcome. Clearly, these more complicated patterns will be difficult to discern with only 4 points in time unless the relation is a very strong one.

Returning now to Figure 31.8, the linear effect of boy academic skill is negative in direction and becomes stronger from grade 4 to 6, but then weaker at grade 7, and then stronger again at grade 8. The relation also appears to become increasingly nonlinear over time. The apparent nonlinear effect is very interesting and could signal a threshold effect or perhaps a nonadditive interaction with some other predictor, but we will not pursue it here. The effect of father SES is negative in direction and is essentially linear and becomes steadily stronger over time. We can get a better look at how the regression weights change over time by plotting them against time for each predictor. Such a plot is shown in Figure 31.9. As can be seen, father SES shows a stronger and more smoothly linear increase over time, which suggests that it is a

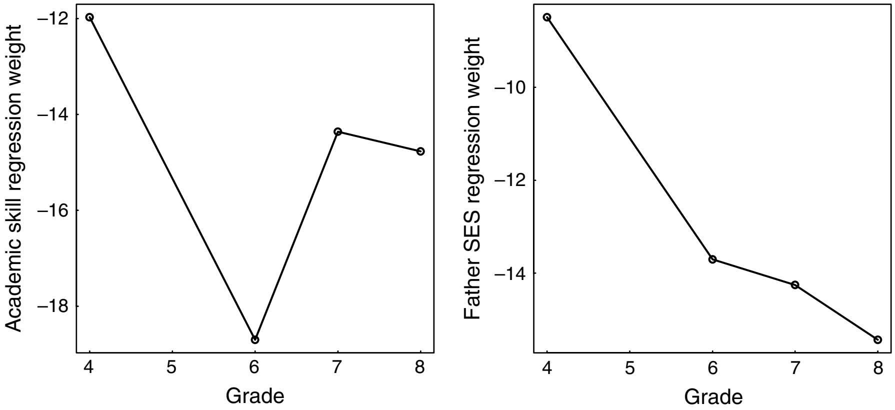


Figure 31.9 Regression weights versus grades for regression of repeated assessments of DPA, grade 4 to 8 on grade 4 academic skill and father SES

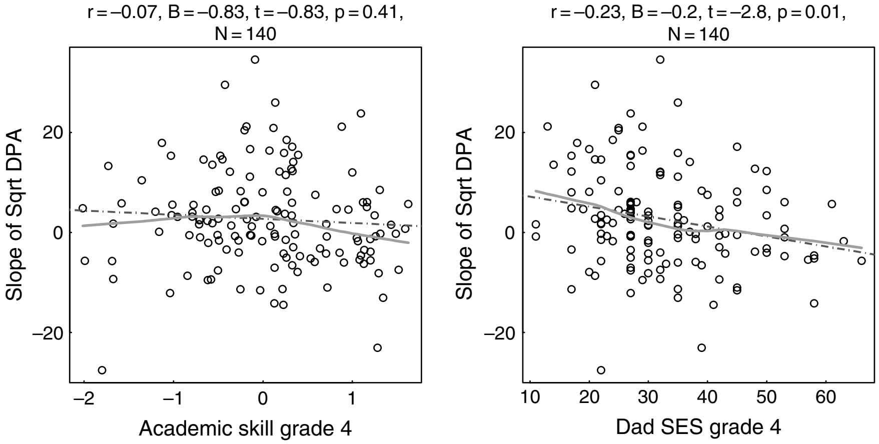


Figure 31.10 Ordinary least-squared slopes for DPA, grade 4 to 8 versus grade 4 predictors, academic skill and father SES

better predictor of linear change in DPA than boy academic skill.

Another useful plot is shown in Figure 31.10. Here, individual growth curves have been fit for each subject by regressing their 4 DPA values on the 4 time values, 0, 2, 3, and 4, in order to get individual slopes. The plot on the right side of the page for father SES shows a significant linear relation with DPA slopes (the test statistics are in the top margin). The plot on the left shows that the boy's academic skill in grade 4 does not predict linear growth in DPA. In Figure 31.11, grade 9 academic skill is plotted against all potential predictors. The bottom right plot looks at the boy's academic skill in grade 9 as a function of linear growth in DPA

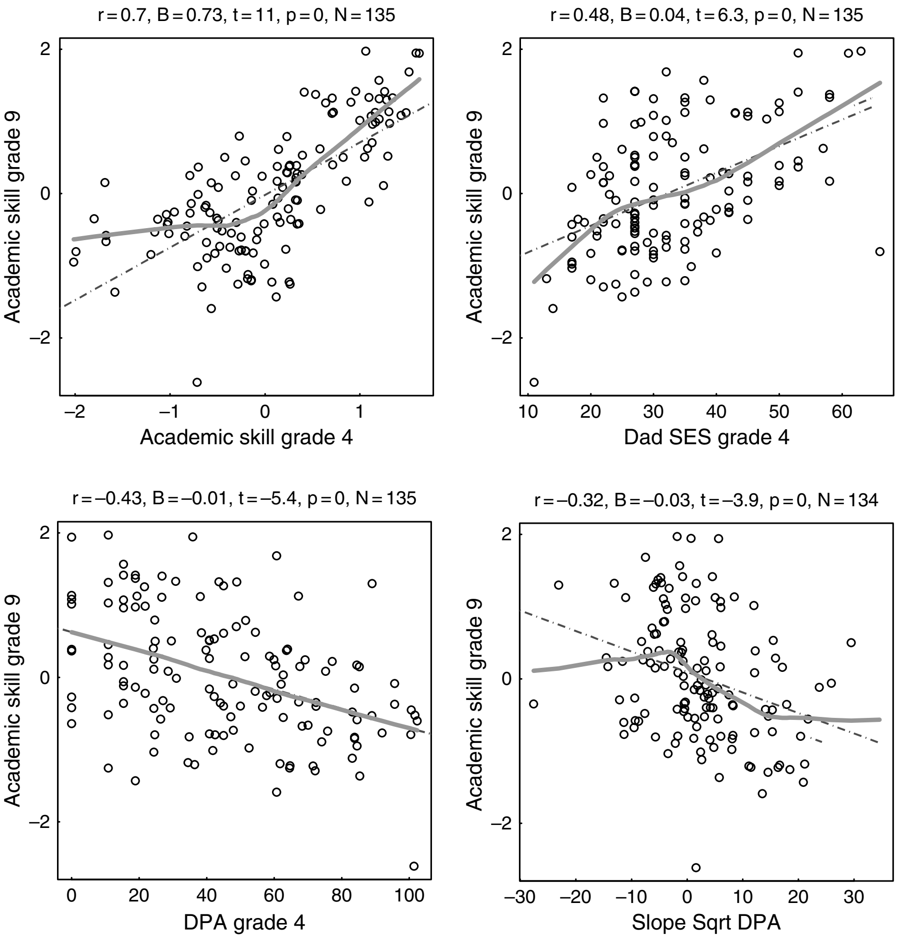


Figure 31.11 Grade 9 academic skill versus grade 4 predictors and ordinary least-squared slopes for DPA, grade 4 to 8

from grade 4 to 8. This plot looks highly nonlinear on both ends of the DPA slope distribution. The nonlinearity in the high end is due to 4 cases and may not be very replicable. The nonlinearity in the low end is more substantial and appears to involve almost one-third of the sample and begin around a DPA slope of zero. The plots on the bottom left and top right for grade 4 DPA and father SES respectively indicate significant linear relations. The plot on the top left for grade 4 academic skill suggests a nonlinear relation on the low end of the grade 4 distribution. The apparent nonlinear relations are interesting and deserve more attention as potential challenges to existing theory if replicable, but we will not pursue them here.

The linear growth model of Figure 31.7 was fit with father SES as the predictor. The results are shown below in Table 31.3. The model has an excellent fit, the chi-square is 4.63 with 11

Table 31.3 Parameter estimates, standard errors (SE) and critical ratios (Est./SE) for DPA growth models with grade 4 father SES (FSES4)

	Estimates	*SE*	*Est./SE*
DPA intercept on father SES 4	−0.79	0.20	−4.01
DPA slope on father SES 4	−0.16	0.06	−2.82
Intercepts			
DPA intercept	71.90	6.79	10.60
DPA slope	7.07	1.94	3.65
Residual variances			
DPA intercept	413.51	69.39	5.96
DPA slope	16.51	5.84	2.83
DPA4-DPA8	356.73	27.58	12.93
Chi-square	4.63		
DF	11.00		
P value	0.95		
CFI	1.00		
TLI	1.02		
RMSEA 90% CI (lower, estimate, upper)	0.00	0.00	0.01

degrees of freedom which generates a p value of .95. The CRI, TLI and RMSEA are also indicative of an excellent fit. The effect of father SES on both intercept and slope factors is highly significant, (z = −4.01 and z = −2.82, respectively) and accounts for 16% of the variance for both the intercept and slope factor. The fact that father SES predicts the intercept and slope factors of DPA means that boys with low SES fathers tend to both start higher in 4th grade on DPA and grow faster from 4th to 8th grade on DPA than boys with high SES fathers. The regression intercepts for the intercept and slope factors are not really interesting because they represent mean levels for boys with father SES scores of zero, which is beyond the range of the observed data in the OYS.

The last step in our introduction to growth curve analysis will be to examine the impact of growth in DPA on future outcomes, in particular academic skill in grade 9, one year after the growth interval. Here, we attempt to assess the consequences of a particular pattern of growth in DPA. Does a high rate of growth on DPA during middle school have a negative impact on academic skill during the first year of high school, over and above initial status of DPA? If so, then it immediately raises the question of whether father SES might also have an impact on academic skill since we have seen that father SES has a significant impact on growth rates of DPA. We can examine whether or not the father SES effect is mediated through growth in DPA or if it has a direct effect on academic skill in grade 9.

First, consider the model shown in Figure 31.4 and suppose we add another y variable, y5, academic skill in grade 9. The new model is shown in Figure 31.12. Notice that the path diagram has been altered compared to Figure 31.4 and in particular it has been simplified by eliminating the circles for the ζs and ε5. As models become more complicated, it usually becomes necessary to start leaving out parts of the path diagram in order to clearly communicate the essential aspects of the model. The intercept and slope factor loadings on y5 are now freely estimated regression weights since academic skill is a conceptually distinct outcome of growth and not part of the DPA series. In addition, unlike y1 to y4, y5 has a nonzero regression intercept indicated by $\nu5$.

The first question is, are these loadings identifiable? The structural equations for the model are

$$\begin{aligned}
y1_i &= \eta1_i + 0\eta2_i + \varepsilon1_i \\
y2_i &= \eta1_i + 1\eta2_i + \varepsilon2_i \\
y3_i &= \eta1_i + 2\eta2_i + \varepsilon3_i \qquad (21)\\
y4_i &= \eta1_i + 3\eta2_i + \varepsilon4_i \\
y5_i &= \lambda_{51}\eta1_i + \lambda_{52}\eta2_i + \varepsilon5_i + \nu_5.
\end{aligned}$$

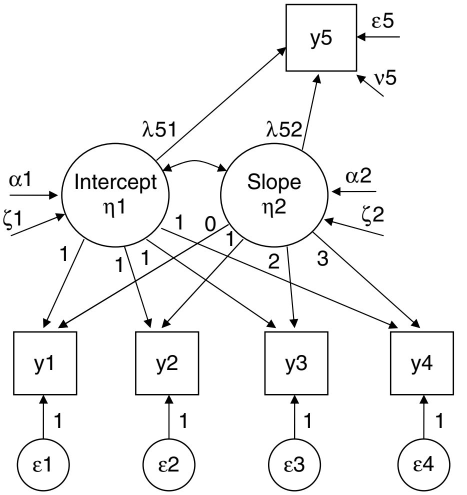


Figure 31.12 Path diagram for linear growth curve model with intercept and slope predicting a distal outcome

The covariance of y5 with y1 is

$$\begin{aligned}
&\operatorname{cov}(y1, y5)\\
&\quad = \operatorname{cov}(\eta 1 + \varepsilon 1, \lambda_{51}\eta 1 + \lambda_{52}\eta 2 + \varepsilon 5 + \nu_5)\\
&\quad = \operatorname{cov}(\eta 1, \lambda_{51}\eta 1) + \operatorname{cov}(\eta 1, \lambda_{52}\eta 2)\\
&\qquad + \operatorname{cov}(\eta 1, \varepsilon 5) + \operatorname{cov}(\varepsilon 1, \lambda_{51}\eta 1)\\
&\qquad + \operatorname{cov}(\varepsilon 1, \lambda_{52}\eta 2) + \operatorname{cov}(\varepsilon 1, \varepsilon 5)\\
&\quad = \lambda_{51}\operatorname{var}(\eta 1) + \lambda_{52}\operatorname{cov}(\eta 1, \eta 2). \qquad (22)
\end{aligned}$$

The covariance of y5 with y2 is

$$\begin{aligned}
&\operatorname{cov}(y2, y5)\\
&\quad = \operatorname{cov}(\eta 1 + \eta 2 + \varepsilon 2, \lambda_{51}\eta 1 + \lambda_{52}\eta 2 + \varepsilon 5 + \nu_5)\\
&\quad = \operatorname{cov}(\eta 1, \lambda_{51}\eta 1) + \operatorname{cov}(\eta 1, \lambda_{52}\eta 2)\\
&\qquad + \operatorname{cov}(\eta 1, \varepsilon 5) + \operatorname{cov}(\varepsilon 2, \lambda_{51}\eta 1)\\
&\qquad + \operatorname{cov}(\varepsilon 2, \lambda_{52}\eta 2) + \operatorname{cov}(\varepsilon 2, \varepsilon 5)\\
&\qquad + \operatorname{cov}(\eta 2, \lambda_{51}\eta 1) + \operatorname{cov}(\eta 2, \lambda_{52}\eta 2)\\
&\qquad + \operatorname{cov}(\eta 2, \varepsilon 5) \qquad (23)\\
&\quad = \lambda_{51}\operatorname{var}(\eta 1) + \lambda_{52}\operatorname{cov}(\eta 1, \eta 2)\\
&\qquad + \lambda_{51}\operatorname{cov}(\eta 2, \eta 1) + \lambda_{52}\operatorname{var}(\eta 2)\\
&\quad = \lambda_{51}[\operatorname{var}(\eta 1) + \operatorname{cov}(\eta 2, \eta 1)]\\
&\qquad + \lambda_{52}[\operatorname{cov}(\eta 1, \eta 2) + \operatorname{var}(\eta 2)].
\end{aligned}$$

We already know from equations (6) to (10) that the variances of η1 and η2 and the covariance are identified so equations (22) and (23) appear to be two independent linear equations in the two unknowns, λ51 and λ52. By applying the determinant test[9] with a11 = var(η1),

[9] The determinant test is fairly easy and useful to use for two linear equations with two unknowns. The two equations will be independent and have a unique solution if the determinant test is not zero. Suppose we have two unknowns, x and y, and two equations for x and y

$$\begin{aligned}
&a11x + a12y = c1\\
&a21x + a22y = c2
\end{aligned}$$

The determinant is a11 a22 − a21 a12 and so long as the determinant is not zero, the equations have a unique solution. For LGC, the coefficients (the a's) will be elements from either the observed mean vector, the observed covariance matrix, or previously identified model parameters. As noted in the main text, the determinant test is zero if the Slope variance is zero and as also previously noted, if our running example model includes the covariance between the Intercept and Slope, the Slope variance is only marginally significant. Being marginally significant means that the Slope variance is close to zero, probabilistically speaking, which means the model is close to being unidentified. When a model is close to being unidentified, the parameters involved frequently have highly inflated standard errors, which indicates that the data do not very precisely estimate the parameters in question and this does indeed happen to the effects of the Intercept and Slope on grade 9 academic skill when the Intercept-Slope covariance is included in the model. Both have highly inflated standard errors and neither is significant until the Intercept-Slope covariance is removed from the model as shown in Table 31.4.

a12 = cov(η1,η2), a21 = var(η1) + cov(η1,η2), and a22 = cov(η1,η2) + var(η2) we find

$$\begin{aligned} a11a22 - a12a21 &= \mathrm{var}(\eta1)[\mathrm{cov}(\eta1,\eta2)+\mathrm{var}(\eta2)] \\ &\quad -\mathrm{cov}(\eta1,\eta2)[\mathrm{var}(\eta1)+\mathrm{cov}(\eta1,\eta2)] \\ &= \mathrm{var}(\eta1)\mathrm{cov}(\eta1,\eta2)+\mathrm{var}(\eta1)\mathrm{var}(\eta2) \\ &\quad -\mathrm{cov}(\eta1,\eta2)\mathrm{var}(\eta1)-\mathrm{cov}(\eta1,\eta2)^2 \\ &= \mathrm{var}(\eta1)\mathrm{var}(\eta2)-\mathrm{cov}(\eta1,\eta2)^2 \\ &= \mathrm{var}(\eta1)\mathrm{var}(\eta2)-\mathrm{cov}(\eta1,\eta2)^2 \\ &\quad \times\mathrm{var}(\eta1)\mathrm{var}(\eta2) \\ &= \mathrm{var}(\eta1)\mathrm{var}(\eta2)[1-\mathrm{cor}(\eta1,\eta2)^2]. \qquad (24) \end{aligned}$$

Equation (24) is only zero when η1 and η2 are perfectly correlated or when either η1 or η2 have a variance equal to zero; so long as neither of these things is true, λ51 and λ52 are identified.

To show that $\nu5$ is identified, the mean of y5 is

$$\begin{aligned} E(y5) &= E(\lambda_{51}\eta1+\lambda_{52}\eta2+\varepsilon5+\nu5) \\ &= \lambda_{51}E(\eta1)+\lambda_{52}E(\eta2)+\nu5 \qquad (25) \\ &= \lambda_{51}\alpha1+\lambda_{52}\alpha2+\nu5. \end{aligned}$$

But, the only unknown in the entire equation is $\nu5$ since we have already shown that the λ's and the α's are identified so $\nu5$ is identified. The last step is to identify the variance of ε5 which is easily done from the variance of y5

$$\begin{aligned} \mathrm{var}(y5) &= \mathrm{var}(\lambda_{51}\eta1+\lambda_{52}\eta2+\varepsilon5+\nu5) \\ &= \mathrm{var}(\lambda_{51}\eta1)+\mathrm{var}(\lambda_{52}\eta2)+\mathrm{var}(\varepsilon5) \\ &\quad +2\mathrm{cov}(\lambda_{51}\eta1,\lambda_{52}\eta2)+2\mathrm{cov}(\lambda_{51}\eta1,\varepsilon5) \\ &\quad +2\mathrm{cov}(\lambda_{52}\eta2,\varepsilon5) \\ &= \lambda_{51}^2\mathrm{var}(\eta1)+\lambda_{52}^2\mathrm{var}(\eta2)+\mathrm{var}(\varepsilon5) \\ &\quad +2\lambda_{51}\lambda_{52}\mathrm{cov}(\eta1,\eta2). \qquad (26) \end{aligned}$$

The only unknown in the entire equation is the variance of ε5 since all other model parameters are identified so it too is identified. Thus, the model with a future outcome predicted by the latent growth factors is completely identified and, in fact, overidentified. In addition, we showed that the model with a predictor of growth could be identified from just the growth model itself. Thus, the model with both early predictors of growth and future consequences of growth is completely identified. If the early predictors of growth also have direct effects on the future consequences, these effects will also be identifiable because this part of the model is just observed variable regression.

Returning now to our empirical example, Table 31.4 below shows results from the growth model with grade 9 academic skill included as future consequence of growth. The effects of the intercept and slope of DPA on academic skill are strongly significant and jointly account for

Table 31.4 Parameter estimates, standard errors (SE) and critical ratios (Est./SE) for DPA growth models with grade 9 academic skill

	Estimates	*SE*	*Est./SE*
Academic skill 9 on DPA intercept	−0.02	0.00	−7.06
Academic skill 9 on DPA slope	−0.09	0.02	−4.15
Means			
DPA intercept	46.27	2.42	19.13
DPA slope	1.96	0.67	2.92
Intercepts			
Academic skill 9	1.16	0.14	8.42
Variances			
DPA intercept	533.93	85.36	6.26
DPA slope	22.89	6.75	3.39
Residual variances			
DPA4-DPA8	350.95	27.27	12.87
Academic skill 9	0.31	0.05	6.02
Chi-square	9.01		
DF	11.00		
P value	0.62		
CFI	1.00		
TLI	1.01		
RMSEA 90% CI (lower, estimate, upper)	0.00	0.00	0.08

58% of the variance of academic skill. High scores on either initial status in grade 4 or growth rate from grade 4 to 8 of DPA, or both, substantially lower grade 9 academic skill.

Finally, we combine both the early predictors and future consequences in one overall model, which is shown in Figure 31.13. Results are shown in Table 31.5. The first four lines show the effects of the growth factors and the grade 4 predictors on grade 9 academic skill. The DPA slope factor and grade 4 academic skill have the strongest effects and both standardized effects (−.44 and .48 respectively) are about equal in magnitude. The effect of the DPA intercept factor is just barely significant at the .05 level (standardized effect = −.21) and the effect for father SES is nowhere close. The stability effect of early academic skill on later academic skill is not surprising but the fact that both DPA intercept and slope predict academic skill at grade 9, and in fact, the standardized DPA slope effect (−.44) is larger than the DPA intercept effect (−.21) highlights the importance of considering change over time on DPA. The predictors jointly account for 75% of the variance of grade 9 academic skill. Both father SES and the boy's academic skill at grade 4 have significant effects on the intercept of DPA although the father SES effect is marginal. Only father SES has

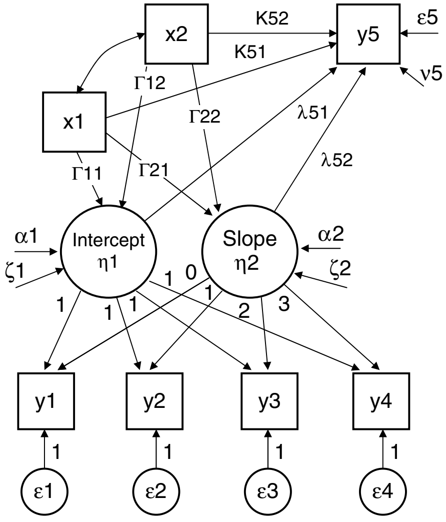


Figure 31.13 Path diagram for linear growth curve model with two pre-existing predictors of intercept and slope and pre-existing predictors, intercept and slope predicting a distal outcome

Table 31.5 Parameter estimates, standard errors (SE) and critical ratios (Est./SE) for DPA growth models with grade 4 academic skill, father SES and grade 9 academic skill

	Estimates	*SE*	*Est./SE*
DPA intercept on academic skill 4	−14.16	2.58	−5.49
DPA intercept on father SES 4	−0.44	0.19	−2.33
DPA slope on academic skill 4	−0.13	0.82	−0.16
DPA slope on father SES 4	−0.16	0.06	−2.63
Academic skill 9 on DPA intercept	−0.01	0.00	−2.13
Academic skill 9 on DPA slope	−0.09	0.03	−3.39
Academic skill 9 on academic skill 4	0.50	0.09	5.90
Academic skill 9 on father SES 4	0.01	0.01	0.74
Intercepts			
Academic skill 9	0.37	0.28	1.32
DPA intercept	61.13	6.44	9.49
DPA slope	7.09	2.04	3.48
Residual variances			
DPA4-DPA8	355.42	27.14	13.10
Academic skill 9	0.19	0.05	4.07
DPA intercept	285.77	54.56	5.24
DPA slope	17.01	5.58	3.05
Chi-square	11.75		
DF	15.00		
P value	0.70		
CFI	1.00		
TLI	1.01		
RMSEA 90% CI (lower, estimate, upper)	0.00	0.00	0.06

a significant effect on the slope of DPA and academic skill does not. The father SES effect is consistent with results in Table 31.2; including grade 4 academic skill in the prediction of the DPA slope does nothing to change those results. There is a significant .34 correlation between father SES and academic skill at grade 4.

The model in Table 31.5 also has indirect effects to consider. Recent work indicates that for indirect effects the sampling distributions for test statistics are not normally distributed as was commonly assumed, especially in samples that are not large (MacKinnon et al., 2002). To compensate for non-normality and asymmetry we use a 95% bias corrected confidence interval generated by a bootstrapping approach available in Mplus for the grade 4 predictors. Father SES has significant indirect effects through both the DPA intercept (lower limit = 0.001, estimate = 0.004, upper limit = 0.011) and slope (lower limit = 0.005, estimate = 0.013, upper limit = 0.034) on grade 9 academic skill. Academic skill at grade 4 has a significant indirect effect through the DPA intercept (lower limit = 0.001, estimate = 0.115, upper limit = 0.236) and a non-significant effect through the slope (lower limit = −0.137, estimate = 0.011, upper limit = 0.193) on academic skill at grade 9.

The model in Table 31.5 has substantive implications for efforts to enhance the scholastic success for boys in high school. It suggests that early efforts to enhance academic skill in elementary school will have direct effects on later academic skill and will also tend to reduce deviant peer affiliation in elementary school which in turn will increase academic skill in high school. Early efforts to reduce deviant peer affiliation will also tend to pay off in terms of higher achievement in high school. A high priority should also be placed on preventing growth in deviant peer affiliation during middle school as this has a substantial direct effect on high school achievement. On a more speculative note, increasing academic skill for boys in high school may also have an intergenerational effect. If increasing academic skill for the boy in high school leads to higher SES when he becomes an adult, this would tend to reduce deviant peer affiliation in his own boys, assuming he marries and has boys. This in turn would lead to greater academic achievement and possibly higher SES.

Author notes

Correspondence by regular mail concerning this chapter should be sent to Mike Stoolmiller, 137 $^1/_2$ W Washington, Marquette, MI, 49855 or by email to mikes@oslc.org. Support for this work comes from Department of Education grant H324×010014.

References

Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd edn. London: Arnold.

Kirk, R. E. (1982). *Experimental Design*, 2nd edn. New York: Brooks/Cole.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38: 963–974.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods,* 7(1): 83–104.

Muthen, L. K. and Muthen, B. O. (2004). *Mplus User's Guide*, 3rd edn. Los Angeles, CA: Muthen & Muthen.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effect Models in S and S-Plus*. New York: Springer.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models*. Thousand Oaks, CA: Sage.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7: 147–177.

Stoolmiller, M. (1994). Antisocial behavior, delinquent peer association, and unsupervised wandering for boys: Growth and change from childhood to early adolescence. *Multivariate Behavioral Research*, 29: 263–288.

Stoolmiller, M., Duncan, T., Bank, L. and Patterson, G. R. (1993). Some problems and solutions in the study of change: Significant patterns in client resistance. *Journal of Consulting and Clinical Psychology*, 61: 920–928.

Chapter 32

Multilevel growth curve analysis for quantitative outcomes

Douglas A. Luke

Multilevel growth curve modeling is a powerful and flexible statistical technique which can be used to model longitudinal data. The primary purposes of growth curve modeling are to describe the form and structure of change in a quantitative dependent variable over time, and to explore the interindividual and intraindividual predictors of this change. Growth curve modeling is a type of multilevel modeling, based on a mixed effects statistical model, which treats multiple observations as nested within individuals. Growth curve modeling has numerous statistical advantages for analyzing longitudinal data. In particular, it can handle missing data and longitudinal designs where observations occur at different time points across individuals. Growth curve models can be fit with any statistical software that includes mixed effects or multilevel modeling procedures. Multilevel growth curve modeling is one of the most powerful and flexible ways to analyze longitudinal data.

1 Introduction

Growth curve modeling is a flexible and powerful way to analyze longitudinal data. The term "multilevel growth curve" recognizes the fact that growth curve modeling is a type of multilevel model where observations are nested within individual cases. The use of the term growth curve arises out of psychology, where this type of multilevel modeling was first used to describe developmental growth of a variety of psychological characteristics (Bryk and Raudenbush, 1987; McArdle and Nesselroade, 2002). However, as we shall see, growth curve modeling can be applied to any form of longitudinal data where the interest is in change—no formal conception of a "growth" process is required.

Multilevel growth curve modeling uses a mixed effects general linear modeling approach to estimate the statistical model. Growth curve models can also be estimated using a latent construct approach via structural equation modeling (SEM) software. This SEM approach is not covered in this chapter. Interested readers can see the excellent introduction to latent growth curve models by Terry Duncan and his colleagues (1999). See also Stoolmiller, Chapter 31, in this volume.

Multilevel growth curve analysis has a number of important strengths. First, it allows flexible statistical modeling that can more closely match the underlying longitudinal theoretical framework. In particular, multilevel growth curve modeling can help disentangle questions about interindividual predictors (e.g., Do children who go to pre-school show quicker mastery of reading skills in elementary school compared to children who do not go to preschool?) from intraindividual predictors (e.g., Does receipt of positive feedback speed up the acquisition of particular reading skills for individual children?).

The primary purpose of this chapter is to introduce growth curve modeling techniques in an applied way so that the interested reader can see their potential for longitudinal data analysis. In the next section some basic considerations about research design and data management in growth curve modeling are discussed. Following this, longitudinal data from the National Longitudinal Survey of Youth are used to illustrate the basic steps in building and evaluating a growth curve model. Finally, the chapter concludes with a short appendix covering software that can be used to fit growth curve models.

2 Research design and data management

Growth curve models are based on longitudinal data from longitudinal research designs. Longitudinal data are made up of observations of one or more dependent and independent variables which are measured on the same individuals at multiple points in time. Technically, prepost data that are obtained at two different time points are longitudinal. However, no real longitudinal research questions can be addressed with such data. Questions about the form of change over time, in particular, can only be answered with data that are measured at three or more time points (Singer and Willett, 2003).

Longitudinal data may be obtained from either experimental or observational studies. For example, a clinical trial study of the effects of an educational campaign designed to promote screening for prostate cancer would collect longitudinal data from participants over time after enrolling in the study. The primary hypothesis would be that participants receiving the new educational materials would show higher rates of screening over time than participants in a control condition. Longitudinal observational studies are also extremely common. For example, using health surveys of adolescents, investigators could track substance use patterns over time. One longitudinal hypothesis could be that students who transfer schools may show steeper increases in substance use over time than students who remain in the same school. From the perspective of the statistical analyst, there is no difference between experimental and observational longitudinal data. The primary difference is in the interpretation of the results—e.g., much stronger claims for causality may be made for experimental longitudinal data than from observational data.

2.1 Data management

Data management for growth curve and other types of longitudinal data analysis can become somewhat complicated. However, all longitudinal datasets will have certain core features. First, longitudinal datasets will have five basic types of variables: an ID variable, one or more longitudinal dependent variables, one or more variables containing time information, time-varying predictors, and time invariant predictors. A dataset used for growth modeling will always have at least the first two types, but the presence of the different types of predictor variables will depend on the study design and research questions.

Although longitudinal data are often initially collected and stored in different data files, eventually the data will be brought together for analysis. There are two common formats for storing longitudinal data, illustrated in Table 32.1. In the “wide” data format, each record in the database is a separate individual. Multiple observations on the same individual are stored in different variables (e.g., weight1, weight2, etc.) in the same case. However, most multilevel software packages will expect to see longitudinal data in a different format, where data are stored in one observation per record. In this "tall" format each observation gets its own record, and any longitudinal data are stored in one variable (e.g., weight). Most general-purpose statistical packages provide routines that can relatively easily restructure the data from one format to the other. Notice that in the observation record format, the multilevel structure of the longitudinal data is apparent: multiple observations are nested within individuals (see below).

2.2 Introduction to the NLSY97 dataset

The data used to provide examples for this chapter are taken from the National Longitudinal Survey of Youth 1997 Cohort (NLSY97). The NLSY97 is part of a series of surveys funded by the US Bureau of Labor Statistics and designed to gather longitudinal data on the labor market experiences of US youth and adults. The NLSY97 examines the transition from school to work for a nationally representative sample of youth who were born from 1980 to 1984. The youths were ages 12 to 17 during the first wave of data collection. 8984 participants were interviewed in 1997, and annual interviews were conducted for the next seven years. The sample size for round 7 was 7756, and the overall retention rate was 86.3%. The NLSY97 collected information on a wide variety of educational, work, and health areas. With the large sample size, number of variables, and up to seven time points for each participant, the NLSY97 is an ideal data source for exploring growth curve modeling. See http://www.bls.gov/nls for more information.

For this chapter, data were extracted and downloaded from the complete seven-year NLSY97 public dataset. We will be focusing on developing growth models for two dependent variables: BMI and Total Substance Use Days. BMI is the body mass index and is an important risk factor for a wide variety of health conditions related to obesity. BMI was not measured directly in the NLSY97, but is based on

Table 32.1 Comparison of the "individual record" (wide) and "observation record" (tall) data structures

Individual record structure							
ID	*Gender*	*Age1*	*Weight1*	*Age2*	*Weight2*	*Age3*	*Weight3*
001	M	12	125	13	129	14	137
002	F	12	101	13	103	14	108

Observation record structure				
ID	*Time*	*Gender*	*Age*	*Weight*
001	1	M	12	125
001	2	M	13	129
001	3	M	14	137
002	1	F	12	101
002	2	F	13	103
002	3	F	14	108

self-reported measures of height and weight. The formula for BMI is

$$BMI = \left(\frac{Weight\ in\ pounds}{(Height\ in\ inches) \times (Height\ in\ inches)}\right) \times 703$$

The NLSY97 asked youth to report the number of days in the past 30 days that they had used alcohol, marijuana, or smoked cigarettes. We combined these three measures to form a total substance use risk variable, called Substance Use Days, that can range from 0 to 90. The higher the number, the more often the youth is reporting using substances in the past month. We will use hierarchical linear modeling to examine how each of these variables change over time as youths age, and we will also explore how certain covariates predict interindividual differences in change patterns over time. The covariates include individual characteristics such as gender and race, as well as one important time-varying predictor, transition to a new school.

3 Building the multilevel growth model

3.1 Framing a growth curve model as a multilevel model

In a traditional regression model, variability of the dependent variable is accounted for either by the predictor variables, or else put into an undifferentiated individual error term. In a multilevel statistical model, as the name suggests, we are able to partition variability across multiple levels. So, for example, if we want to understand reading achievement by students in multiple classrooms, using a multilevel model we can account for variability that exists between students (level 1) and also variability between classrooms (level 2). That is, students are nested in classrooms, and we can build statistical models that reflect that reality.

Growth curve models are simply a special type of multilevel model. Here, multiple observations across time are nested within individuals. As we stated earlier, a principal advantage of multilevel modeling is its ability to account for nonindependence of observations due to nesting. So, just as we might expect students in the same classroom to be more similar to one another than would be expected by chance (thus violating the traditional independence assumption), we certainly expect multiple observations of the same person to be more alike. Growth curve modeling using hierarchical linear models can appropriately account for this nonindependence of observations across time.

The following system of equations shows a basic growth curve model as a multilevel model:

$$Y_{ti} = \beta_{0i} + \beta_{1i} T_{ti} + \varepsilon_{ti}$$
$$\beta_{0i} = \gamma_{00} + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + u_{1i}$$

Here we are modeling some dependent variable Y, measured at time t on individual i. The first line of the model can be considered as the level 1 portion of the model, and looks similar to a typical multiple regression model. The only level 1 predictor included in this basic growth curve model is T, which is the time variable. For this reason this model is sometimes called an *unconditional linear growth curve model*. It is unconditional in that there are no predictors, other than the time variable. It is linear in that β_{1i} only captures the linear relationship between time and the dependent variable.

The most important difference between the above model and a traditional multiple regression model can be seen by considering the presence of the i subscripts in the level 1 portion of the model. Both the intercept and slope betas have i subscripts, indicating that we are allowing the intercepts and slopes to vary across individuals. That is, each individual in the dataset is allowed to have his or her own growth curve! For this reason, we often call the level 1 part of the growth curve model the *intraindividual* part of the model.

In a multilevel model, the parameters in the first level of the model become outcomes in the second level of the model. In the unconditional linear growth model, the intercept for a particular individual (β_{0i}) is predicted by the grand mean of all the individual intercepts (γ_{00}) plus the variability of the individual intercepts around the grand mean (u_{0i}). Similarly, the slope for a particular individual (β_{1i}) is predicted by the grand mean of all the individual slopes (γ_{10}) plus the variability of the individual slopes around the grand mean (u_{1i}). The level 2 part of the model is called the *interindividual* part of the model, because it can be used to model predictors of change between individuals.

Instead of using a system of equations to specify the multilevel model, we can substitute the level 2 parts of the model into the level 1 equation. After substituting and rearranging the terms, we get the following:

$$Y_{ti} = \underbrace{[\gamma_{00} + \gamma_{10} T_{ti}]}_{fixed} + \underbrace{[u_{0i} + u_{1i} T_{ti} + \varepsilon_{ti}]}_{random}$$

This single prediction equation form of the multilevel model is called the *mixed effects model*, because it shows how the model is based on both fixed effects (the gammas γ) and random effects (the variance components ε and u). Although it is harder to discern the multilevel structure of the model when it is in this form, it more clearly states what components are actually being modeled. This form of the model also closely corresponds to the output of the various multilevel modeling software packages. It is advantageous to be able to construct and interpret multilevel models using both types of equations. Fortunately, hierarchical linear modeling software such as HLM allows you to see the models in both forms.

The unconditional linear growth model, presented above, is only one of an innumerable set of possible growth curve models. This simple model can be extended by adding predictors at either of the levels of the model, as well as by making decisions about which random effects to include in the model. It is difficult at this point to know how to make these decisions. So in the next two sections we will look at how to extend the model by considering two fundamental questions about growth curve models. First, how can we describe the form of change over time? Second, what factors influence intra- and interindividual patterns of change?

3.2 Describing the form of change

A starting point for most growth curve models is to describe the form or shape of change in the dependent variable of interest. The purpose of this first step may simply be descriptive, or it might be to address a specific scientific question (e.g., "Does the increase in BMI during adolescence follow a quadratic form?").

Before jumping into fitting and testing specific multilevel growth models, it is advisable to spend some time thinking theoretically about expected patterns of change. This can help guide the often complicated process of model selection. In addition, it is always a good idea to examine the data to see what the individual raw growth curves look like. Figure 32.1 presents 40 randomly chosen plots of the raw growth curves of BMI from the NLSY97 data. It is apparent that BMI levels vary substantially between youth. However, it appears that for many of the youth, BMI tends to go up as they get older. Although the patterns are not consistent across youth, it appears that changes in BMI may not proceed in a simple linear fashion. So we might want to examine more complicated growth models that describe nonlinear change.

Figure 32.1 also reveals that a number of the participants do not have measurements for all seven time points. In fact, just in this random sample we see one person with only one measurement, and a couple of people with

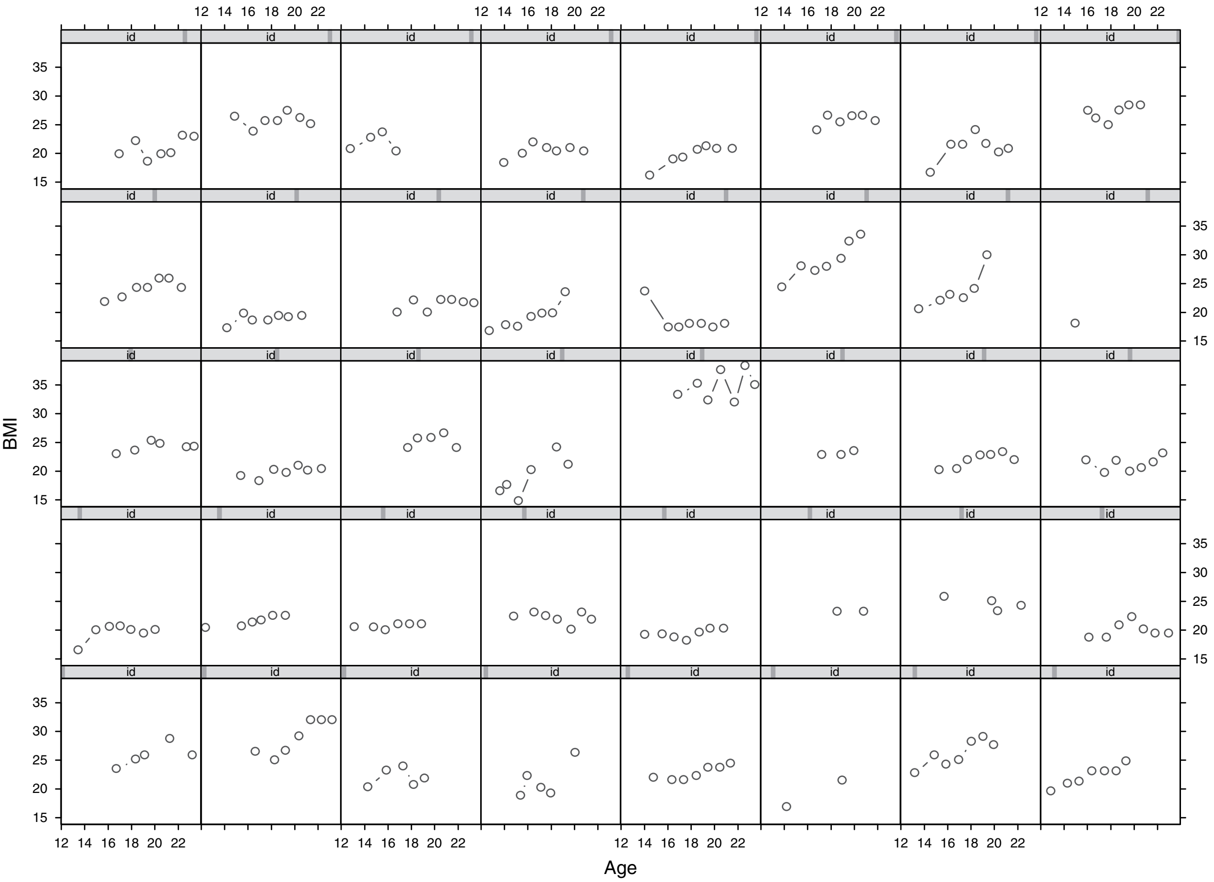


Figure 32.1 Individual growth curves of BMI by age for 40 random cases

only two measurements. Very close examination of the figure also reveals that the period between measurements varies across individuals. That is, although youth are interviewed *on average* once a year in the NLSY97, the actual time between interviews can be quite a bit shorter or longer for particular individuals. These two common aspects of real world longitudinal datasets, varying number of measurements and varying time between measurements, pose severe or even fatal challenges for traditional longitudinal statistical approaches such as repeated measures ANOVA. However, multilevel modeling can handle this type of “messy” data without any problem. This is, in fact, one of the primary reasons that multilevel modeling is now a preferred analytic approach for growth curve models.

Unless you have a very specific hypothesis about the form of the change, a reasonable approach to model building is to start simple, and then build more complex models. For this reason, we start by fitting the above unconditional linear growth model to BMI:

$$BMI_{ti} = \beta_{0i} + \beta_{1i}\,(Age12)_{ti} + \varepsilon_{ti}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}$$

In this model we are predicting BMI scores for individuals across time. Time is measured using the age of the youth at the time of the interview after subtracting 12. (We will explain the reason for this below.) Both the intercept and time slope are allowed to vary across individuals. With this model we will be able to see how much BMI goes up or down as youths age. Given our examination of the raw growth curves, we might expect the time slope to be positive.

The results of this model are presented under Model 1 in Table 32.2. This table summarizes much of the useful information from a growth model. The top of the table presents estimates of the fixed effects. The coefficients are estimates of the two gammas. The average intercept across all individuals is 20.63. This means that the expected BMI for any individual when the age variable is 0 is approximately 21. This helps explain why we subtracted 12 from the age at interview variable. If we used the raw age at interview, then the estimate of the intercept would be interpreted as the expected BMI score when a person was age 0 (i.e., a newborn). This, of course, is not a useful or interpretable estimate. By subtracting 12 from each age at interview score, we get a new interpretation of the intercept—the expected BMI value at age 12. We picked age 12 because this is approximately the youngest age for which there are data in the NLSY97 dataset. This is an example of *centering* a predictor variable. There has been a lot written about centering variables in multilevel models (see, e.g., Paccagnella, 2006). Although the topic can get quite complicated, the most important reason to center predictor variables is to produce fixed effects estimates that are more interpretable than would otherwise be the case.

The linear fixed effect estimate of 0.54 tells us that for each additional year of age, we would expect BMI to increase by about half a point for each individual. This suggests that during the teen and young adult years, youth are becoming more overweight as they age. Along with the coefficient estimates, their associated standard errors, *t*-tests, and *p*-values are presented. Hypothesis testing of the individual fixed effects parameters can thus be done using traditional methods.

The middle rows of Table 32.2 present the random effects part of the model. These are presented in the form of variance components, and can be thought of as unmodeled variability. The variance component of 16.64 tells us that there is a large amount of variability of individual BMI scores around the average starting point of 20.63. This confirms what we saw in Figure 32.1, where we saw some people with quite low BMI scores, and others with high BMI scores. The much smaller linear variance component of 0.23 suggests that there is much less variability of the slopes across individuals. One way to view this is that there is much more variability left over to model (with predictor variables) of the *level* of BMI, than there is of the *slope* of BMI on age. Finally, the level 1 variance component estimate of 3.31 suggests that there is a moderate amount of intraindividual variability. This suggests that the individual observations may be bouncing around the linear regression line. This could be due to instability of the BMI measurements. Another possibility is simply that the simple linear growth model is not a good fit with the data.

In addition to the variance component estimates, some multilevel modeling software packages will produce statistical tests of these components. Here we see the chi-square tests and associated *p*-values produced by HLM. However, these statistical tests should be viewed with caution. First, variance components are bounded at 0, so their distributions are not normal. Second, it is not clear exactly what the meaning of a significant variance component should be—after all, we generally expect variances to be nonzero. Rather than focusing on the *p*-values of the variance components, it is usually more fruitful to interpret their

Table 32.2 Three growth models for change of BMI

Fixed effects	*Model 1 – Linear*				*Model 2 – Quadratic*				*Model 3 – Cubic*			
	Coef.	*SE*	*t-ratio*	*p*	*Coef.*	*SE*	*t-ratio*	*p*	*Coef.*	*SE*	*t-ratio*	*p*
Intercept (γ_{00})	20.63	0.050	413.7	0.000	20.172	0.067	302.7	0.000	19.77	0.088	225.5	0.000
Linear (γ_{10})	0.54	0.006	84.6	0.000	0.722	0.020	36.2	0.000	1.00	0.046	21.4	0.000
Quadratic (γ_{20})					−0.016	0.002	−9.7	0.000	−0.07	0.008	−8.2	0.000
Cubic (γ_{30})									0.003	0.000	6.4	0.000
Random effects	*Variance component*		χ^2	*p*	*Variance component*		χ^2	*p*	*Variance component*		χ^2	*p*
Intercept (u_{0i})	16.64		39260	0.000	16.54		14437	0.000	16.46		14381	0.000
Linear (u_{1i})	0.23		26028	0.000	0.93		10439	0.000	0.91		10384	0.000
Quadratic (u_{2i})					0.006		10403	0.000	0.006		10360	0.000
Cubic (u_{3i})									—[a]		—	—
Level 1 (ε_{ti})	3.31				3.10				3.10			
Model fit												
Deviance	260895.1				260257.6				260209.0			
Parameters	6				10				11			
AIC	260907.1				260277.7				260231.0			
BIC	260960.6				260366.7				260329.0			

[a] Cubic effect set to fixed to avoid convergence problems.

sizes rather than their significance (Pinheiro and Bates, 2000).

3.3 Assessing nonlinear change patterns

Model 1 tells us that BMI increases with age, but there is still a lot of variability both across and within individuals. Figure 32.1 suggested that increases in BMI may not be strictly linear in form, so the next step is to build models that will assess the extent to which the form of the change of BMI is nonlinear. There are a number of ways of building such curvilinear models. One of the simpler approaches is to build a polynomial growth model by adding quadratic, cubic, quartic terms, and so on, to the base linear model. For a dataset with *k* time points, in principle *k*-1 polynomial terms can be fit. However, in practice growth models are rarely built that go beyond cubic or quartic components. First, in most areas of the social and health sciences theories are not rich enough to suggest or explain such high-level polynomial models. Second, in many real-world datasets quadratic or cubic models explain most of the intraindividual variability, and it is unusual to have underlying variability that requires more complicated models.

To fit polynomial models, you simply add the appropriate time variable raised to the degree of the polynomial. So, a quadratic model would include Time, and Time-squared. A cubic model would include Time, Time-squared, and Time-cubed, and so on. The following equation is for a quadratic polynomial model of change of BMI:

$$BMI_{ti} = \beta_{0i} + \beta_{1i}\,(Age12)_{ti} + \beta_{2i}\,(Age12)^2_{ti} + \varepsilon_{ti}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}$$

$$\beta_{2i} = \gamma_{20} + u_{2i}$$

Models 2 and 3 listed in Table 32.2 present the results of fitting a quadratic and cubic model, respectively. For both models, the individual coefficients are highly significant, suggesting that a curvilinear model is more appropriate than a simple linear model. In polynomial models, the meaning of the coefficients changes. For example, in a quadratic model, the linear coefficient for time no longer represents a constant change rate. Instead, it now represents the instantaneous rate of change at the point that time = 0. The quadratic coefficient tells us how fast the instantaneous rate of change itself changes. This can be thought of as a *curvature* parameter (Singer and Willett, 2003). Instead of interpreting each coefficient individually in a polynomial model, it is often more informative to plot the prediction curves for the model based on the fitted coefficients. Figure 32.2 presents the prediction curves for BMI for the three models presented in Table 32.2.

The figure shows that BMI increases steadily as youths age. The curvilinear models both suggest that BMI rises faster at early ages, from about 12 to 15. The quadratic model suggests that after about the age of 22 the increase in BMI starts slowing down. To see how these models fit the data at the two age extremes, individual marks were added to the plot that represent the raw average BMI scores for that age group. These marks suggest that the quadratic and cubic models fit the data pretty well for the

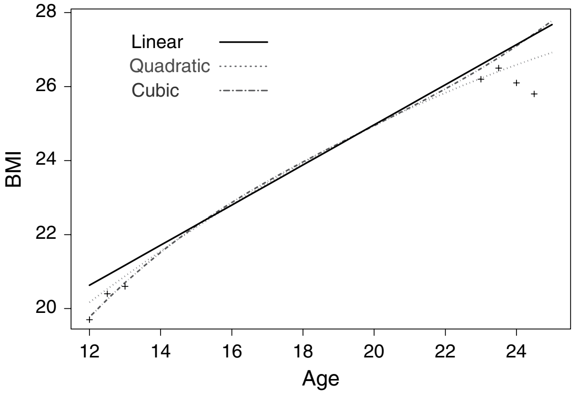


Figure 32.2 BMI prediction curves for linear, quadratic, and cubic growth curve models

Table 32.3 Three growth models for Substance Use Days

Fixed effects	*Model 1 – Linear*				*Model 2 – Quadratic*				*Model 3 – Cubic*			
	Coef.	*SE*	*t-ratio*	*p*	*Coef.*	*SE*	*t-ratio*	*p*	*Coef.*	*SE*	*t-ratio*	*p*
Intercept (γ_{00})	0.030	0.161	0.2	0.852	−2.846	0.174	−16.3	0.000	−0.145	0.214	−0.7	0.497
Linear (γ_{10})	1.655	0.030	55.4	0.000	2.777	0.083	33.44	0.000	0.677	0.173	3.9	0.000
Quadratic (γ_{20})					−0.092	0.007	−13.1	0.000	0.326	0.035	9.3	0.000
Cubic (γ_{30})									−0.024	0.002	−11.9	0.000
Random effects	*Variance component*		χ^2	*p*	*Variance component*		χ^2	*p*	*Variance component*		χ^2	*p*
Intercept (u_{0i})	80.75		15080	0.000	27.25		6486	>0.500	27.11		6426	>0.500
Linear (u_{1i})	4.26		21205	0.000	26.51		9218	0.000	26.35		9147	0.000
Quadratic (u_{2i})					0.162		9579	0.000	0.161		9508	0.000
Cubic (u_{3i})									—[a]		—	—
Level 1 (ε_{ti})	100.24				90.02				89.74			
Model fit												
Deviance	431258.6				428258.7				428116.0			
Parameters	6				10				11			
AIC	431270.6				428278.7				428138.0			
BIC	421324.0				428367.8				428236.0			

[a] Cubic effect set to fixed to avoid convergence problems.

early ages, while the quadratic model does a better job for young adults. (The means for the youngest, age 12, and oldest, age 24.5, groups are based on very small numbers of cases, so they should be interpreted with caution.) An interesting thing to note about these models is that from about the ages of 15 to 22, all three models would lead to virtually the same predicted values.

Table 32.3 and Figure 32.3 present the results for the same set of growth curve models applied to the Substance Use Days dependent variable. As youths age, we see that the number of substance use days also goes up. Again, we find that there are significant curvilinear components to the change over time. Figure 32.3 shows that a linear model does a particularly poor job of predicting substance use for the oldest members of the sample. Conversely, the quadratic model gives impossible predictions for kids aged 12 or 13. The cubic model may do the best job, and it describes a type of S-curve that seems reasonable for this dependent variable. When kids are very young, substance use is near zero and changes slowly. During the teen years substance use increases significantly, but as adulthood approaches, substance use appears to level off.

3.4 Model diagnostics, fit and selection

In addition to examining individual parameters and their associated *p*-values, it is usual to examine model diagnostics to see how well the fitted model matches its underlying assumptions, and then to examine various fit indices to see how well the overall model fits the data.

Diagnostics for growth curve models of quantitative dependent variables are very similar to those examined for multilevel modeling. Two common assumptions that can be easily checked are normality of errors (residuals) and homoscedasticity. Figure 32.4 shows two diagnostic plots on a random 2% sample (~1000 cases) of the quadratic BMI growth model (Model 2 from Table 32.2). The Q-Q plot on the left side tells us that although the residuals are symmetric, they are more kurtotic (higher central peak, smaller tails) than we would expect with independent and normally distributed errors. This suggests that our model may require a more complex covariance structure than we assumed. (Our model was fit assuming a compound symmetry covariance matrix. For details on how to fit growth curve models with other covariance structures, see Singer and Willett, 2003). On the right side we plot the residuals against the fitted (predicted) BMI values. This plot shows no evidence of a fan shape, and strongly suggests that this model does not have problems with heteroscedasticity. Luke (2004) provides more examples of how to use graphical exploration of residuals to examine the assumptions of multilevel models.

Growth curve models for quantitative dependent variables are typically fitted using some form of maximum-likelihood estimation (Laird, 1978). Simply stated, this type of estimation works by maximizing a likelihood function that assesses the joint probability of simultaneously observing all of the sample data, assuming a certain set of fixed and random effects. An important product of the estimation process is a number obtained by multiplying the natural log of the likelihood by –2. This number, sometimes called the *deviance* or designated as

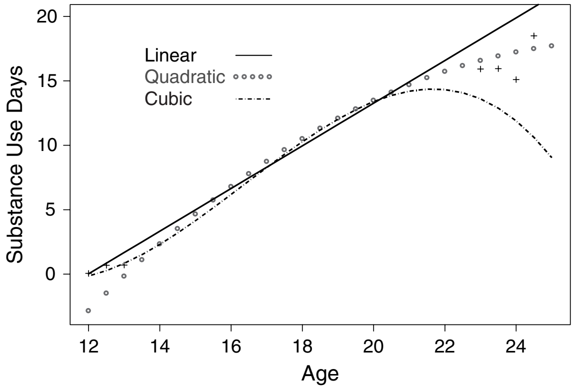


Figure 32.3 Substance Use Days prediction curves for linear, quadratic, and cubic growth curve models

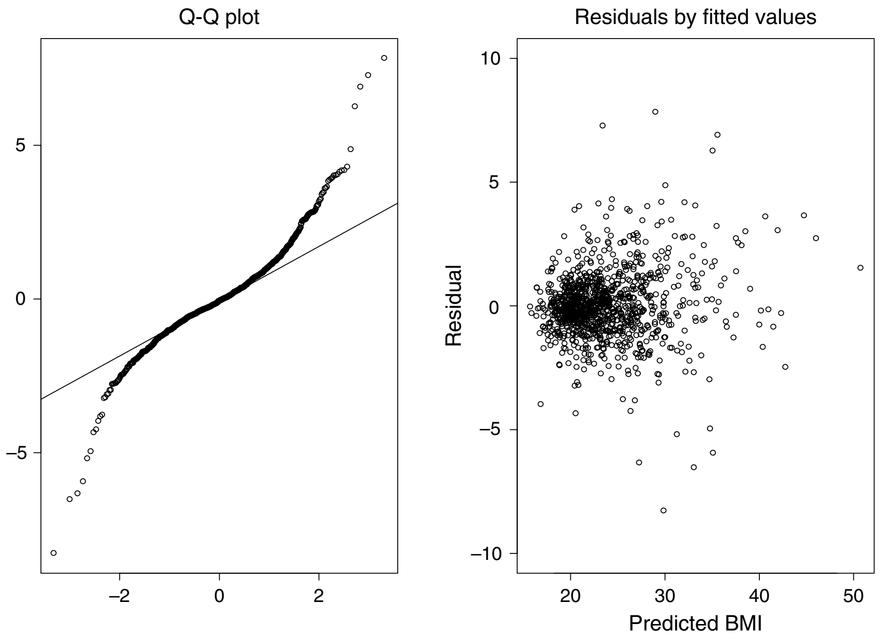


Figure 32.4 Two diagnostic plots for the quadratic BMI growth curve model

−2LL, is a measure of the discrepancy between the observed data and the fitted model. The deviance for any one model cannot be interpreted directly, but it can be used to compare multiple models to one another.

The model comparison can be done between two models fit to the same data, where one of the models is a subset (has fewer parameters) of the other. The difference of the deviances from each model is distributed as a chi-square statistic with degrees of freedom equal to the difference in the number of parameters estimated in each model. For example, we can compare Model 3 to Model 1 for BMI (Table 32.2) to see if the nonlinear change model is better than the simpler linear change model. The difference between the two deviances is 886.1 (260895.1 − 260209.0); this value is highly significant with $df = 5(11-6)$. This tells us that the more complicated model is a significantly better fit to the data.

One disadvantage of the deviance (−2LL) is that a model fit to the same data with more parameters will always have smaller deviance. This is generally good, because smaller deviance implies a better fit to the data. However, we can always get better fit by adding more predictors. We also want to choose the simplest model that describes the data; i.e., the model with the fewest parameters. Two widely used fit indices have been developed that are based on the deviance, but incorporate penalties for a greater number of parameters: The Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC) (Akaike, 1987; Schwarz, 1978). For both of these indexes, smaller is better. Also, an important advantage of these two criteria is that they can be used to compare two models fit to the same dataset, even if one is not a subset of the other. The AIC and BIC are listed in Tables 32.2 and 32.3 for our change models. In both cases these criteria indicate that the cubic models are better models than the simpler linear models. For more details on how to calculate and use AIC and BIC, see Luke (2004).

3.5 Different choices for scaling time

The concept of time is critical for growth curve models. Therefore, it is also critical to think carefully about how to operationalize, measure, and model time in a growth curve model. Time can be defined many ways for any particular study—and these different conceptions may not all equally represent the theory or research question under consideration, and different operationalizations of time may lead to different model results. For example, consider Table 32.4, which shows four different ways that time may be assessed for the NLSY97 study: the interview wave (from 1 to 7), the actual age at the time of the interview, the age after subtracting 12, and the grade that the student is in. If the investigator is interested in the underlying physiological and cognitive changes that influence body weight, then Age or Age12 might be appropriate conceptions of time. On the other hand, if one wants to understand how the changes in school environment may influence drug use, then perhaps Grade would be a useful operationalization of time. However, it is hard to think of a research question that would be usefully served by using Interview Wave as a measure of time. This is an arbitrary time measurement that is based on the logistics of the study, not a physical or social reality.

Table 32.4 Examples of different definitions of time

ID	*Interview wave*	*Age*	*Age12*	*Grade*
001	1	13	1	7
001	2	15	3	8
001	3	16	4	8
001	4	16	4	9
001	5	18	6	11
001	6	19	7	13
001	7	20	8	14
002	1	15	3	10
002	2	16	4	11
002	3	16	4	12
002	4	18	6	13
002	5	19	7	14
002	6	20	8	15
002	7	21	9	15

Table 32.5 presents the fixed effects results of a linear BMI growth model for three different definitions of time. The results for Model 3 on the right side of the table are for the Age12 variable, and are the same as displayed in Table 32.2. Model 1 presents a growth model where Interview Wave is used for time. The results are similar to Age12, but not identical. They are similar in that on average across all subjects each interview wave is approximately one year apart. Both models show that BMI increases about half a point a year. However, consideration of the AIC and BIC scores shows that the model with Age12 is doing a better job of describing the observed data. This is not surprising, because Age12 provides information not just on the order of the interviews, but also reflects an accurate measure of the actual amount of time that has passed between each interview for each participant.

The difference between Models 2 and 3 is simpler and more subtle. Age is the raw age, while Age12 is raw age minus 12. Subtracting (or adding) a constant from a predictor variable is a way of *centering* the predictor variable. Centering is typically done in one of three ways: (1) by subtracting a meaningful constant, as we have done here with Age12; (2) by subtracting a grand mean; or (3) by subtracting a group mean. This third type of centering is more complicated than the other two, but is relatively uncommon for growth models (where the group is each individual). In growth modeling, centering is typically done by subtracting a constant or grand mean, and this has two advantages. First, centering is typically done so that the interpretation of intercepts is more

Table 32.5 Comparison of three BMI linear growth models with different definitions of time

Fixed effects	*Model 1 – Interview wave*				*Model 2 – Age*				*Model 3 – Age12*			
	Coef.	*SE*	*t-ratio*	*p*	*Coef.*	*SE*	*t-ratio*	*p*	*Coef.*	*SE*	*t-ratio*	*p*
Intercept (γ_{00})	21.72	0.047	463.9	0.000	14.12	0.108	131.1	0.000	20.63	0.050	413.7	0.000
Linear (γ_{10})	0.58	0.007	83.0	0.000	0.54	0.006	84.6	0.000	0.54	0.006	84.6	0.000
Model fit												
Deviance	261356.6				260895.1				260895.1			
Parameters	6				6				6			
AIC	261368.6				260907.1				260907.1			
BIC	261422.0				260960.6				260960.6			

meaningful. Remember that an intercept is the predicted value of a dependent variable when the predictors are all 0. Consider Models 2 and 3 in Table 32.5. The intercept for BMI in Model 2 is 14.12. We interpret this as the predicted value of BMI when a person is 0 years old. This interpretation is not useful—BMIs for infants are not defined or interpretable! If we center age by subtracting 12 from each score, the only change in the model is that the intercept is now 20.63. This is the predicted value for a person who is 12 years old ($\text{Age} - 12 = 0$). This is much more meaningful, because it represents the value for the youngest persons who were actually included in the NLSY97 study. If we had centered age by subtracting the grand mean of age across all of the participants and time points, we would have a different intercept. This grand-mean centered intercept would be interpreted as the expected BMI score for a person who was the average age of all persons in the study. For growth models it is fairly typical to center the time variable by subtracting the time at the first observation. This allows an interpretation of the intercept as the "starting point" of the growth curve.

The second reason that centering is typically done in growth models has to do with problems of multicollinearity. If polynomial transformations of time are used to build nonlinear growth models, the various time predictors (time, time-squared, etc.) are highly intercorrelated, and may lead to convergence problems, especially with smaller datasets. By centering all of the time variables the intercorrelations are reduced, and convergence problems will be less likely.

3.6 Identifying predictors of change

The above presentation has focused on developing growth curve models whose primary goal is to describe the form or shape of change. Typically, however, researchers are also interested in developing models and testing hypotheses that include predictors of change. In addition to the level 1 Time variable, growth curve models can include other types of level 1 (intraindividual) and level 2 (interindividual) covariates or predictors.

3.7 Predictors of interindividual change

In growth curve models, covariates that are constant over time, such as gender or experimental condition, are known as interindividual predictors. These predictors can tell us how change varies across different types of individuals. For our example, we will examine the effects of gender and ethnicity on the change in Substance

Use Days. To start we examine whether gender (Male) and ethnicity (NonWhite) affect the intercept of Substance Use Days (SUD). Both of these predictors are binary, with 1 indicating male or nonwhite, respectively. (The original NLSY97 ethnicity variable was categorical with several ethnicity options. This was recoded to 0=white, 1=nonwhite for this example.) The following equation shows the growth model to be fitted.

$$SUD_{ti} = \beta_{0i} + \beta_{1i}\,(Age12)_{ti} + \varepsilon_{ti}$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\,(Male)_i + \gamma_{02}\,(NonWhite)_i + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}$$

This growth model makes it clear that Male and NonWhite are entered as level 2 predictors for the intercept (β_{01}) of Substance Use Days. Note that this means for this first model that we assume a single linear slope for SUD on Age.

The results of fitting this first predictor model are shown in the left-hand side of Table 32.6. Both the gender and ethnicity predictors are highly significant. The intercept (1.39) is now interpreted as the predicted number of Substance Use Days for a white female age 12. The gender effect (1.17) tells us that 12-year-old males use substances about one day a month more often, and the ethnicity effect (−4.15) tells us that nonwhites are much less likely to use substances when they are young.

However, it may be that the relationship between age and substance use may not be the same across the different gender and ethnic groups. To test this, we can fit a more complex model that allows the linear slope of

Table 32.6 Effects of gender and ethnicity on change in Substance Use Days

Fixed effects	*Model 1 – Intercept effects*				*Model 2 – Slope and intercept effects*			
	Coef.	*SE*	*t-ratio*	*p*	*Coef.*	*SE*	*t-ratio*	*p*
Intercept (γ_{00})	1.392	0.242	5.7	0.000	1.708	0.290	5.9	0.000
Male (γ_{01})	1.170	0.228	5.1	0.000	−1.690	0.321	−5.3	0.000
NonWhite (γ_{02})	−4.150	0.226	−18.4	0.000	−1.722	0.319	−5.4	0.000
Age12 slope (γ_{10})	1.661	0.030	55.6	0.000	1.582	0.050	31.8	0.000
Male (γ_{11})					0.738	0.058	12.6	0.000
NonWhite (γ_{12})					−0.624	0.059	−10.6	0.000
Random effects	*Variance component*		χ^2	*p*	*Variance component*		χ^2	*P*
Intercept (u_{0i})	84.62		15280	0.000	79.53		15000	0.000
Linear Slope (u_{1i})	4.27		21241	0.000	4.03		20699	0.000
Level 1 (ε_{ti})	100.07				100.24			
Model fit								
Deviance	430916.6				430648.0			
Parameters	8				10			

SUD on Age12 to vary by gender and ethnicity. This corresponds to the following growth model:

$$SUD_{ti} = \beta_{0i} + \beta_{1i}\,(Age12)_{ti} + \varepsilon_{ti}$$
$$\beta_{0i} = \gamma_{00} + \gamma_{01}\,(Male)_i + \gamma_{02}\,(NonWhite)_i + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}\,(Male)_i + \gamma_{12}\,(NonWhite)_i + u_{1i}$$

Here we can see that gender and ethnicity are allowed to affect not only the intercept of SUD, but also the slope of SUD on Age. This model can be re-expressed in the mixed effects format as:

$$\begin{aligned} SUD_{ti} = {} & \gamma_{00} + \gamma_{01}\,(Male)_i + \gamma_{02}\,(NW)_i \\ & + \gamma_{10}\,(Age)_{ti} + \gamma_{11}\,(Male)_i\,(Age)_{ti} \\ & + \gamma_{12}\,(NW)_i\,(Age)_{ti} + u_0 + u_1 + \varepsilon_{ti} \end{aligned}$$

Although somewhat complicated, the mixed effects version highlights the fact that by including level 2 predictors of the slope we are actually entering cross-level interactions into the model. For example, γ_{11} will assess the extent to which the slope of SUD on age varies between girls and boys. This is, in effect, an interaction between gender (level 2) and time (level 1). Cross-level interactions are often the effects of most interest to researchers. For example, in longitudinal clinical trials, the test of the effectiveness of an intervention is typically modeled as a cross-level interaction of experimental condition (e.g., experimental vs control groups) by time.

The results of this model can be seen in the right side of Table 32.6. The parameter estimates show that gender and ethnicity are strong predictors of both the intercept and slope of time. To interpret these effects, it again helps to plot the prediction equations (Figure 32.5). Here we see the same basic finding that substance use increases over time. However, this model also shows us that we expect this increase to be the greatest for white males, and the lowest for nonwhite females.

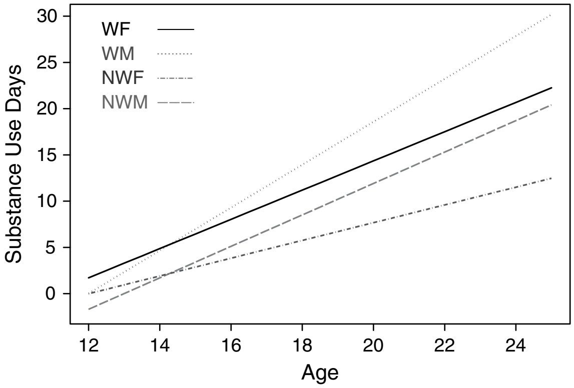


Figure 32.5 Predicted effects of gender and ethnicity on linear change of total Substance Use Days

3.8 Predictors of intraindividual change

There are many times that the predictors of interest in growth curve modeling will change over time themselves. These time-varying covariates can be used to model intraindividual change. Consider, for example, the effects of school transitions on daily substance use among youth. Social scientists have viewed school transitions as times of higher stress for students, as well as opportunities to form new social networks, have more freedom from previous family and peer expectations, and otherwise provide a changed social environment for substance use. It would be reasonable to assume that substance use may look different after a school transition. How could this be modeled?

First, consider Table 32.7, which shows an example data file that could be used in growth curve modeling. Male is a variable denoting gender that is a constant covariate—it does not change over time during the study. College transition, on the other hand, is an indicator variable that is 1 when an interviewee is attending a new college during the time of the interview. This is a time-varying covariate—it can take on different values (although only two values for a binary variable) at different time points.

Table 32.7 Example data file with constant and time-varying covariates

ID	*Interview wave*	*Age*	*Substance Use Days*	*Male*	*College transition*
001	1	13	0	1	0
001	2	15	3	1	0
001	3	16	8	1	0
001	4	16	4	1	0
001	5	18	10	1	1
001	6	19	12	1	0
001	7	20	10	1	0
002	1	15	2	0	0
002	2	16	5	0	0
002	3	16	8	0	0
002	4	18	15	0	1
002	5	19	14	0	0
002	6	20	12	0	1
002	7	21	12	0	0

Using the NLSY97 data, we can examine the effects of college transition on substance use over time using the following model:

$$
\begin{aligned}
SUD_{ti} = \beta_{0i} + \beta_{1i}\,(College)_{ti} + \beta_{2i}\,(Age12)_{ti}\\
+ \beta_{3i}\,(College)\,(Age12)_{ti} + \varepsilon_{ti}
\end{aligned}
$$

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

$$\beta_{1i} = \gamma_{10}$$

$$\beta_{2i} = \gamma_{20} + u_{3i}$$

$$\beta_{3i} = \gamma_{30}$$

The college transition variable is entered into the level 1 part of the growth model, because it can take on different values at different time points (as suggested by the *t* subscript). College appears twice in the level 1 part of the model. The college main effect (β_{1i}) will assess the effects of college transition on the intercept of Substance Use Days. That is, it will allow us to see how much substance use shifts up or down during a year when there is a college transition. The college by age interaction term (β_{3i}), on the other hand, allows us to see if there is a change in the slope of substance use over time after a college transition. This model has no level 2 predictors, and note that only the intercept and Age12 are modeled as random effects. We assume for this example that the effects of college transitions are the same for all individuals.

The results of fitting this model are presented in Table 32.8. Similar to our previous models, we see that youth start out at age 12 using substances approximately 0 days, and that this increases by about 1.6 days of use per year. A transition to college is associated with an upward shift of 4.2 substance use days. However, after a college transition, the upward trend over time has been reduced by .47 days per year. This can be seen more clearly in the prediction graph in Figure 32.6. In this graph we examine the predicted growth curve of substance use days for a person who enters college at age 18. The vertical dashed line represents

Table 32.8 Effects of college transitions on change in Substance Use Days

Fixed effects	*Model 1 – Intercept effects*			
	Coef.	*SE*	*t-ratio*	*p*
Intercept (γ_{00})	−0.007	0.161	−0.0	0.967
College (γ_{10})	4.160	0.923	4.5	0.000
Age12 (γ_{20})	1.649	0.030	54.8	0.000
College X Age12(γ_{30})	−0.471	0.126	−3.7	0.000
Random effects	*Variance component*		χ^2	*p*
Intercept (u_{0i})	80.64		15083	0.000
Linear Slope (u_{1i})	4.26		21222	0.000
Level 1 (ε_{ti})	100.12			
Model fit				
Deviance	431225.1			
Parameters	8			

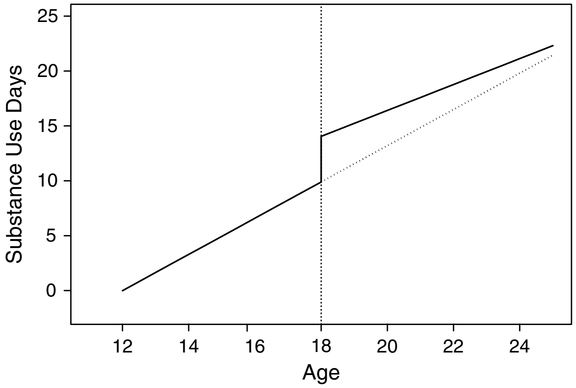


Figure 32.6 Predicted effects of college transition (at age 18) on linear change of Substance Use Days

the transition to college at age 18. At that point we see the sudden shift of substance use days upward to about 14 days. This may reflect the greater access to alcohol that many college students experience. After the transition to college, substance use still increases, but at a slower rate than for youth who do not make a transition to college (represented by the dashed line that continues upward to the right).

4 Conclusion

As we have seen, multilevel growth curve modeling is a flexible tool for analyzing longitudinal data. By viewing longitudinal data as observations nested within individual cases, we can use the power of multilevel modeling to answer questions about patterns and predictors of change. A number of advanced or more technical topics have been passed over or only mentioned briefly in this chapter. In particular, this chapter has focused on the use of growth curve modeling for quantitative dependent variables. Growth curve models can be built for other types of dependent variables, including binary, count, and ordinal variables. For more detailed treatment of these generalized multilevel models, see the relevant sections in Hox (2002) and Snijders and Bosker (1999). Also, the next chapter in this volume deals with multilevel change models for categorical dependent variables.

Software

Users have a large number of good choices for software for fitting growth curve models. Any statistics package that includes mixed effects modeling or multilevel modeling can be used to develop growth curve models of the type discussed in this chapter. Table 32.9 lists the major software packages that are widely known and are powerful enough to develop a wide variety of growth curve models. Users can choose to use specialized software that focuses primarily on multilevel modeling (i.e., HLM or MLwiN), or general-purpose statistical software that includes mixed effects modeling procedures (i.e., R/S-Plus, SAS, SPSS, or Stata). Users new to growth curve models may want to learn these procedures using the specialized software. The interface and documentation of these packages make for a shallower learning curve for growth curve modeling. More experienced users may wish to use the general-purpose software. In particular, the data management and graphical exploration features of R, SPSS, SAS, and Stata cannot be matched by HLM or MLwiN.

The Centre for Multilevel Modelling (sic) maintains a comprehensive list of reviews of software packages for multilevel modeling at: http://www.mlwin.com/softrev/index.html. All of the packages listed in Table 32.9 are reviewed at this site, but some of the reviews are out of date. An extremely useful site for learning about multilevel software is UCLA's statistical computing portal at: http://www.ats.ucla.edu/stat/. For example, all of the data and examples from Singer and Willett's textbook on longitudinal data analysis are presented for each of the six software packages listed in Table 32.9. See http://www.ats.ucla.edu/stat/examples/alda.htm.

Table 32.9 Information about growth curve modeling software

	Specialized multilevel modeling software			
	Version	*Interface*	*Information*	*Core references*
HLM	6.02	Graphical	http://www.ssicentral.com	Raudenbush, et al. (2000)
MLwiN	2.01	Graphical	http://www.mlwin.com	Rasbash, et al. (2000)
	General statistics software			
	Version	*Interface*	*Information*	*Core references*
R/S-Plus – nlme or lme4	R: 2.3.0; S-Plus: 7	Syntax	http://www.r-project.org/ http://www.insightful.com	Pinheiro and Bates (2000)
SAS – Proc MIXED	9.1.3	Syntax	http://www.sas.com	Singer (1998)
SPSS – MIXED	14	Either	http://www.spss.com	SPSS Advanced Models documentation
Stata – gllamm and xtmixed	9	Syntax	http://www.stata.com http://www.gllamm.org	Rabe-Hesketh and Skrondal (2005)

Glossary

AIC Akaike Information Criteria—a parsimony corrected measure of model fit.

BIC Bayesian Information Criteria—a parsimony corrected measure of model fit.

Centering A reparameterization of a predictor by subtracting a grand mean, group mean, or constant.

Deviance This is –2 times the log-likelihood of an estimated model.

Fixed effect This corresponds to the constant effects across persons of a predictor variable in a growth curve model.

Growth curve model A mixed effects model applied to longitudinal data.

Maximum likelihood estimation The most common type of estimation technique used for growth curve models of quantitative dependent variables.

Mixed effects model A statistical model incorporating both fixed and random effects, useful for analyzing grouped and longitudinal data.

Random effect This corresponds to the variance components in a growth curve model. Parameters (slopes and intercepts) that are allowed to vary across persons are random effects.

References

Akaike, H. (1987). Factor analysis and the AIC. *Psychometrica*, 52: 317–332.

Bryk, A. S. and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin,* 101: 147–158.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F. and Alpert, A. (1999). *An Introduction to Latent Variable Growth Curve Modeling.* Mahway, NJ: Lawrence Erlbaum.

Hox, J. (2002). *Multilevel Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution.

Journal of the American Statistical Association, 73: 805–811.

Luke, D. A. (2004). *Multilevel Modeling.* Thousand Oaks, CA: Sage.

McArdle, J. J. and Nesselroade, J. R. (2002). Growth curve analyses in contemporary psychological research. In J. Schinka and W. Velicer (eds), *Comprehensive Handbook of Psychology,* Volume 2: *Research Methods in Psychology*. New York: Pergamon Press.

Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review*, 30: 66–85.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed Effects Models in S and S-Plus*. New York: Springer.

Rabe-Hesketh, S. and Skrondal, A. (2005). *Multilevel and Longitudinal Modeling Using Stata.* College Station, TX: Stata Press.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. and Lewis, T. (2000). *A User's Guide to MLwiN*. London: University of London, Institute of Education.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F. and Congdon, R. (2000). *HLM 5: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood, IL: Scientific Software International.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6: 461–464.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics,* 24: 323–355.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York: Oxford University Press.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advance Multilevel Modeling*. London: Sage.

Chapter 33

Multilevel analysis with categorical outcomes

Scott Menard

In the analysis of intraindividual change using quantitative dependent variables, it is common to apply some form of growth curve analysis, either the latent growth curve model described by Stoolmiller in Chapter 31, or the multilevel growth curve model described by Luke in Chapter 32, both in this volume. With a categorical dependent variable, however, even the definition of “change” is not entirely obvious (as illustrated in Chapter 30 on panel analysis with logistic regression), and “growth” only makes sense for variables measured at the ratio, interval, or at least ordinal level. For a dichotomous dependent variable, one cannot go “up” or “down” on a continuum; one can only change from one value to the other (and back again). For a polytomous (multiple categories) nominal dependent variable, there is no inherent meaning to the “direction” of change. Thus, it is more appropriate to speak of qualitative “change” rather than implicitly quantitative “growth” when dealing with categorical dependent variables. In modeling longitudinal patterns of change, moreover, the latent growth curve modeling approach has not been very extensively or well adapted to the analysis of categorical outcomes. Instead, in terms of both conceptual development and even more in terms of readily available software, models for categorical developmental change are most readily implemented in the multilevel modeling approach, with some of the simpler models being possible using population averaged and random effects models discussed in other chapters (Hilbe and Hardin, Chapter 28; Greenberg, Chapter 17; Finkel, Chapter 29; Menard, Chapter 30).

In this chapter, considerations pertinent to the use of multilevel growth curve models for categorical dependent variables are introduced in the context of analyzing a multilevel change model of marijuana use. We begin with issues in identifying the relationship between the dependent variable and the time dimension used in the model, with particular attention to the use of orthogonal polynomial contrasts. We then examine the general structure of the multilevel model, and its specific application to categorical dependent variables. Following a description of the data to be used in the examples, we then calculate population averaged (marginal) and unit-specific (or subject-specific, or conditional) models for the prevalence of marijuana use, including random as well as fixed effects in the models. (It may be useful here to review the material on marginal and conditional models with categorical dependent variables in Chapter 30.) The focus here is on the multiwave longitudinal multilevel model with a dichotomous outcome. The chapter ends with a brief discussion of extensions of the models presented here to include additional interaction effects, and to the inclusion of polytomous dependent variables.

1 Specifying the relationship between the categorical dependent variable and time

For any model involving the pattern or timing of change in a categorical dependent variable, a useful step in constructing logistic regression change models consists of screening the data to determine the zero-order relationship between the dependent variable and time. For a dichotomous dependent variable, it may be informative to plot the mean of the dependent variable (the mean probability of a response of 1 as opposed to a response of 0) with respect to the time dimension (age or chronological time). Polytomous nominal dependent variables are problematic in this respect, because there is no "increase" or "decrease" in the dependent variable, only change. Here, for manageably short series, a contingency table of the dependent variable with time may be useful for exploring the data, but for longer series (e.g., more than 20 or so cases) this becomes unwieldy. A polytomous ordinal dependent variable can, for this purpose, be treated as an interval variable, and the average (mean or median) rank can be plotted against age or chronological time. Such plots, however, may not by themselves be adequate, particularly when the focus is on change *within* cases. One approach to examining patterns of intraindividual change is to plot the values of the dependent variable along the time dimension for each individual separately, possibly overlaying the plots. For a small number of cases, this may produce distinct and distinguishable patterns, but with over 1000 cases, the plot is likely to be an indistinct blob, possibly obscuring even outlying cases. (For a more extended discussion of issues in graphing intraindividual change across time, see Fitzmaurice, Chapter 13, in this volume).

With large numbers of cases and a focus on change within cases, therefore, initial screening may best be done using a model with a polynomial function of time as a predictor. The degree of the polynomial (the highest power included in the function) needs to be at least one less than the number of time points (period or age) against which the dependent variable is plotted, and as a practical matter should probably be about three less than the number of time points. The model would then be logit(Y) $= \ln[Y/(1-Y)] = \alpha + \beta_1 t + \beta_2 t^2 + \ldots + \beta_{T-3} t^{T-3}$ where t is the time variable (age or period), and there are a total of $t = 1, 2, \ldots, T$ distinct periods for which the model is being estimated. This procedure, although especially informative for explaining intracase change for a large number of cases, can be informative for deciding the function of time to be used, at least initially, in the full model. For dichotomous dependent variables, estimating a bivariate model with age or time as the only predictor can be estimated using an *orthogonal polynomial contrast.* (For a general discussion of the use of different contrasts for categorical predictors in logistic regression analysis, see Menard, 2002.) For a polytomous dependent variable, the same procedure can be followed by first selecting an appropriate contrast, then using that contrast to dichotomize the dependent variable into c−1 dichotomous variables, where c is the number of categories in the dependent variable. The c−1 dichotomous variables will correspond to the c−1 logistic functions to be modeled. For a nominal polytomous dependent variable, an indicator contrast corresponding to the baseline category logit model would be most appropriate; for an ordinal polytomous dependent variable, a contrast corresponding to the type of ordinal logistic regression (e.g., cumulative or continuation ratio logit) should be used. Once again, an orthogonal polynomial contrast for the predictor (age or time) should be used, but there will be c−1 separate logistic regression equations to estimate.

For age (A) or time (T) in their original metrics, the different powers of the polynomial will be highly correlated, e.g., A with A^2 and A^3, producing collinearity among the powers of A and making it difficult or impossible to separate linear (A) from quadratic (A^2) and cubic (A^3) effects. Centering A by subtracting the mean eliminates the correlation between the linear and quadratic effects: $(A-\mu_A)$ should be uncorrelated with $(A-\mu_A)^2$. For higher powers, however, centering is generally not sufficient, and more complex methods of calculating orthogonal polynomials may be required. An orthogonal polynomial contrast can be automatically implemented in some existing software (e.g., SAS or SPSS). The equation being modeled can be expressed as $\text{logit}(Y) = \alpha + \beta_1 f_1(t) + \beta_2 f_2(t^2) + \ldots + \beta_{c-1} f_{c-1}(t^{c-1})$ where t is again the time dimension (age or chronological time) and f(t) is a transformation designed to make the contrast involving time orthogonal.

By examining the statistical significance of the categorical time dimension variable as a whole, it is possible to see whether age or time has any statistically significant *bivariate* influence on the dependent variable, but given the possibility of a suppressor effect, statistical nonsignificance of the time dimension in the bivariate analysis is not a sufficient basis for eliminating the time dimension from the full model. A slightly different approach may be followed, however, in examining the statistical significance of the specific powers of the time dimension. As a general observation, it is rare in both the social and the physical sciences to find a relationship that requires higher than a cubic power between a predictor and an outcome. For powers higher than 4, there is a danger that the model may be overfitted, and that the additional variation being explained by the higher powers of the polynomial may be random rather than systematic variation in the dependent variable. For this reason, it is generally reasonable to eliminate all powers higher than the last statistically significant power of age or time from the model for further estimation.

If it is the case that there are statistically nonsignificant coefficients for powers of t *lower* than the highest power of t for which there is a statistically significant coefficient, there are three options: (1) stepwise elimination of statistically nonsignificant powers of t using backward elimination, regardless of whether the nonsignificant powers of t thus eliminated are lower than the highest power of t for which a statistically significant coefficient is obtained; (2) hierarchical elimination, in which all powers of t *lower* than the highest power of t for which there is a statistically significant coefficient are retained; or (3) forward inclusion, stopping when the next power of t is statistically nonsignificant. The danger of using option (3) is the same as the danger of using forward stepwise inclusion generally, misspecification by the omission of effects that would be discovered as statistically significant in a full model or a reduced model using backward elimination. Option (2) is probably the safest option in the sense of avoiding misspecification, but it runs the risk of inefficiency by the inclusion of unnecessary parameters in the model. Option (1) seems like a reasonable balance between the two. In general, hierarchical elimination (option 2) seems to be the preferred option in practice, but there is generally little reason to expect the function of time expressed as a higher order polynomial to be hierarchical in nature. Bear in mind, too, that the polynomial form of the function of time may be only an approximation to another function of time, not a representation of the true function of time, and a higher order polynomial may reflect this.

Further insight on this may be attained when other predictors are added into the model, if coefficients for higher order functions of t are no longer statistically significant in the presence of statistical control for other predictors. The initial results regarding the powers of t that are statistically significant, however, do give us a starting point for estimating the model with a particular function of time. It may not be the case, however, that higher powers of time are necessary in the full model. Instead, the apparent influence of the higher powers of time may instead reflect the interaction between the time dimension and one or more *time-varying covariates,* predictors whose values, like that of the dependent variable, can change over time. This can be tested by comparing models (1) having *both* the higher order powers of the time dimension *plus* the time-covariate interactions in the model, with models having (2) the time-covariate interactions but not the higher powers of the time dimension, and (3) the higher order powers of the time dimension but not the time-covariate interactions. This may be relevant in models of either historical or developmental change, but seems to be more typical of developmental change.

2 Multilevel logistic regression models for repeated measures data

Multilevel change models of intraindividual (or more generally intracase) change are discussed in Bijleveld et al. (1998, Ch. 5), Raudenbush and Bryk (2002, Ch. 6), and Snijders and Bosker (1999, Ch. 12). Raudenbush et al. (2000) and Snijders and Bosker (1999) include chapters on analysis of categorical (dichotomous, nominal, ordinal, and count) dependent variables. The basic model for multilevel analysis of longitudinal data involves two levels, the individual or case level (level 2), with data that describe characteristics of the case that do not vary over time, and the observation level (level 1), with data on repeated measurements of time-varying individual characteristics, including the dependent variable. A simple descriptive change model would include no level 2 predictors, and only a measure of time or age (or both) as a predictor in the level 1 model. In this case the effect of time on the dependent variable is said to be fixed (as opposed to random, i.e., variable). More complex models could include more complex functions of time (e.g., quadratic or cubic polynomials) and additional time-invariant covariates at level 2 plus time-varying covariates at level 1.

The *level 1 equation* in this context represents the *repeated observations* nested within individuals or cases, and has the form

$$\eta_{tj} = \text{logit}(Y_{tj}) = \beta_{0j} + \beta_{1j}X_{1tj} + \beta_{2j}X_{2tj} + \cdots + \beta_{kj}X_{ktj} + r_{tj} \qquad (1)$$

where the subscript $t = 1, 2, \ldots, T$ refers to the measurement times, $k = 1, 2, \ldots, K$ refers to the predictors $X_1, X_2, \ldots, X_K$, and $J = 1, 2, \ldots, J$ refers to the specific cases (typically individuals) for which the parameters $\beta_0, \beta_1, \ldots, \beta_J$ are calculated; and β_{0j} is the intercept (instead of α_j) to simplify the multilevel notation. The dependent variable η is a function of Y in a generalized linear model, and our specific concern is with $\eta_{tj} = \text{logit}(Y_{tj})$ or a parallel transformation for the nominal and ordinal polytomous logistic regression models. The term r_{tj} represents a random effect (essentially random error) at level 1. The predictors $X_1, X_2, \ldots, X_K$ are *time-varying covariates,* predictors which like the dependent variable represent repeated observations or measurements nested within the cases. The level 2 predictors $W_1, W_2, \ldots, W_Q$ are *time-constant* covariates, stable characteristics of the cases on which the repeated measurements (of at least the dependent variable) are taken over time. At level 2 (the *case* or, most often, the *individual* level), we can model the level 1 coefficients as a function of an intercept (an individual or case mean value) and the $q = 1, 2, \ldots, Q$ time-constant covariates:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}W_{k1} + \gamma_{k2}W_{k2} + \cdots + \gamma_{kQ}W_{kQ} + u_{kj} \qquad (2)$$

where u_{kj} is the level 2 random effect. Combining the level 1 and level 2 models,

$$\eta_{tj} = (\gamma_{00} + u_{0j}) + [\Sigma_k \Sigma_q (\gamma_{kq}W_{kq} + u_{kj})]X_{ktj} + r_{tj} \quad (3)$$

where the respective level 1 β coefficients have been replaced by their level 2 equations. Possible models include pure fixed effect models (all $u_{kj} = 0$ by definition), *random intercept* models with fixed coefficients ($u_{0j} \neq 0$ but all other $u_{kj} = 0$ by definition), and more broadly *random coefficient* (random slope plus random intercept) models (random components u_{kj} are included in at least some of the β coefficients). If the only random coefficient is the random intercept, it is possible to estimate the model using something other than multilevel software, e.g., Stata **xtlogit** for a dichotomous dependent variable. More complex models with additional random coefficients can be estimated using dedicated multilevel modeling software that permits the use of categorical dependent variables, such as HLM.

Unlike multilevel growth curve models for interval/ratio dependent variables, the use of multilevel change modeling for categorical dependent variables will typically involve less of a concern with variance components, and particularly with the allocation of variance in the dependent variable between the level 1 and level 2 components of the model, than is the usual practice using multilevel modeling. Instead, the focus will generally be on (a) how well we can explain the variation in the dependent variable, and (b) the statistical and substantive significance of different predictors as predictive or explanatory variables. In logistic regression analysis more generally, the focus is most appropriately on explained variation, as measured by the likelihood ratio coefficient of determination R_L^2 (Menard, 2000); but in the context of multilevel modeling, estimation often involves approaches other than maximum likelihood (see, e.g., Raudenbush and Bryk, 2002) and hence the appropriate maximum likelihood statistics for calculating R_L^2 are typically not available. In this case, the squared correlation between the observed and predicted values of the dependent variable, R_O^2, may be the best we can do for a measure of explained variation. For individual predictors, particularly when different predictors are measured on different scales, and especially for predictors measured on arbitrary metrics, the fully standardized logistic regression coefficient $b^* = (b)(s_X)(R)/s_{\text{logit}(\hat{Y})}$ as defined in Menard (2004) can be used to indicate substantive significance. In this formula, b^* is the standardized coefficient, b is the unstandardized coefficient, s_X is the standard deviation of the predictor, R is the correlation between the observed (zero or one) and predicted (probability ranging from zero to one) values of the outcome, and $s_{\text{logit}(\hat{Y})}$ is the standard deviation of the predicted values of Y.

3 Multilevel logistic regression for prevalence of marijuana use

To illustrate the application of the unit-specific and population averaged multilevel models in logistic regression for a dichotomous dependent variable, we turn once again (as in Chapter 30) to the prevalence of marijuana use as our dependent variable, with exposure, belief, gender, and ethnicity as predictors, using data from the National Youth Survey (NYS), a multiwave longitudinal study of a self-weighting national household probability sample of 1725 individuals who were 11–17 years old when first interviewed in 1977, and who were last interviewed in 2002. As described previously in Chapter 30, the dependent variable is marijuana use, or more specifically change in the prevalence (yes or no) of marijuana use. The predictors are (1) exposure to delinquent friends, a scale indicating how many of one's friends have engaged in nine different types of delinquent behavior ranging from assault to theft to illicit and underage substance use and drug sales, plus whether they have encouraged the respondent to do anything against the law; (2) belief that it is wrong to violate the law, a scale indicating how wrong the respondent thinks it is to engage in any of nine types of behavior (the same as the first nine items in the exposure to delinquency scale), (3) age, measured as years since birth, (4) gender, coded 0=female, 1=male, and (5) ethnicity, with white non-Hispanic respondents as the reference category and two other categories, African-American and Other.

Exposure and belief are both time-varying covariates on which we have repeated measures across the different ages and periods, and thus will be included in the level 1 model. Gender and ethnicity are time-invariant covariates, unchanging characteristics of the individuals within which the observations are nested, and will be included in the level 2 model. Age varies over time, but does so identically for every case (everyone ages at the same rate). Theoretically, marijuana use should be positively associated with exposure to delinquent friends and negatively associated with belief that it is wrong to violate the law. Age, ethnicity, and gender are included as demographic controls, but associations of age, gender and ethnicity with marijuana use have been found in past research. For a more detailed description of the sample, the variables, and the theoretical basis for the models tested here, see Elliott et al. (1989). Here, since we want to model the relationship between marijuana use and age, we begin by examining the mean age-specific prevalence of marijuana use plus the relationship between prevalence of marijuana use and age with an orthogonal polynomial contrast for age.

Figure 33.1 shows the pattern of the mean prevalence of marijuana use with age. (A plot of the mean of the logit of marijuana use with age has a similar shape but different numerical values.) It appears from the plot that marijuana use peaks around ages 17–19, then declines (a little irregularly) up to age 33, the oldest age for which data are available in this NYS dataset. The pattern suggests that there is at least a quadratic relationship involving $(\text{Age})^2$, and possibly higher powers of age. Analysis of the relationship between marijuana use and the orthogonal polynomial contrast of age indicated that the linear coefficient of Age, or the coefficient of $(\text{Age})^1$, was not statistically significant, but the coefficients for the second, third, and fourth powers of Age were statistically significant as predictors of marijuana use. These results were used to construct the predictors Agep1, Agep2, Agep3, and Agep4 (*Age* to the *power 1*, *Age* to the *power 2*, . . . , *Age* to the *power 4*) for inclusion in the analysis of marijuana use over the portion of the life span included in the NYS data. A word of caution: the fourth order polynomial function of age appears to be appropriate for the age range included in the sample, but it would generally be inappropriate to try to generalize the results beyond the ages included in the sample. In particular, projecting the fourth order polynomial function of age below age 11 or above age 33 to predict prevalence of marijuana use at earlier or later ages would be inappropriate, and such projections, taken far enough beyond the age ranges for which they were estimated, typically produce nonsensical results.

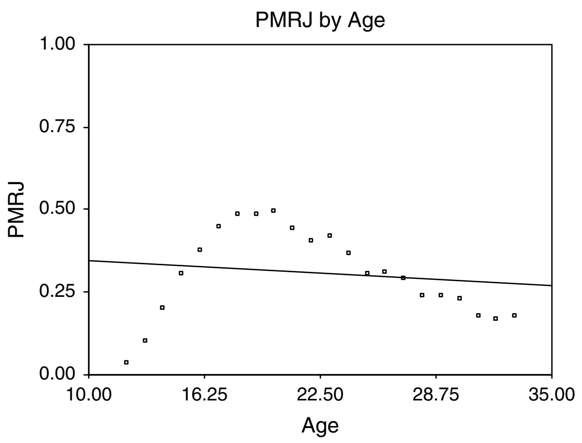


Figure 33.1 Mean prevalence of marijuana use by age

We also need to contend with a feature of the National Youth Survey (NYS) data that may occur in other datasets as well: unequal measurement intervals. For the data used here, measurements were taken for the years 1976, 1977, 1978, 1979, 1980, 1983, 1986, 1989, and 1992.

Also, marijuana use and exposure are measured for the interval spanning the entire year for which measurement occurs, but belief is measured for a point in time at the end of that measurement period. If we lag both exposure and belief as predictors of marijuana use (as one might in a two or three wave panel model), we have a lag of three years between belief and exposure as predictors and marijuana use as an outcome. Possible solutions to this problem include (1) ignoring the length of the measurement period, risking misspecification and underestimation of the impact of exposure and belief in the later waves of the model; (2) ignoring the incorrect temporal order of belief relative to exposure and marijuana use, risking overestimation of the relationship of belief to marijuana use and calling into question the causal direction of the relationship; or (3) using a one-year lag for each wave, and where necessary imputing the values of the lagged predictors, belief and possibly exposure as well. There are several possible implementations of this last approach. The simplest would be linear interpolation, calculating the change per year and then subtracting that amount from the values for exposure and belief as measured for 1983, 1986, 1989, and 1992. A more complex alternative would be to use information about the distribution of exposure and belief over time, plus information about *all* of the waves of data (instead of just using the information for the current and previous wave) to estimate the missing values. Rather than focus on the issues involved in missing value imputation, linear interpolation is used here to produce estimates of the lagged values of the predictors exposure and belief.

4 The population averaged model for prevalence of marijuana use

In a population averaged model, the concern is with rates or averages in the population, with how much an increase in *average* exposure or *average* belief would have on the *average* prevalence of marijuana use in the population. This model would typically be more appropriate for (1) historical change, when all of the cases in the population experienced the same historical influences, and (2) broadly-based population level as opposed to individual-level interventions, when all of the cases received the same intervention. The population averaged or marginal model includes within-case interdependence of the observations by averaging effects across all cases. Table 33.1 presents the results of calculating a population averaged model for the prevalence of marijuana use with gender, ethnicity, exposure, and belief (both lagged to be measured temporally prior to the measurement of marijuana use), and the orthogonal fourth power polynomial function of age. Not shown here are tests for collinearity, which indicated that collinearity among the predictors was not a problem. The R_O^2 statistic for the model indicates a moderate level of explained variance ($R_O^2 = .28, p = .000$). Also included in this row is the number of degrees of freedom for level 2 (individual respondents) and level 1 (observations nested within respondents), used in calculating the statistical significance of the Student's t statistics (not to be confused with the variable t for the time dimension) for the coefficients.

The dependent variable (logit marijuana use) is listed along with the predictors in the first column in Table 33.1. The second column shows the standardized logistic regression coefficient (b^*), the third presents the unstandardized logistic regression coefficient (b), and the fourth contains the standard error of the unstandardized logistic regression coefficient. The fifth column shows a Student's t statistic based on the estimated unstandardized coefficient and its standard error, from which the p value is found based on the number of degrees of freedom, in this instance 1672 for the intercept, gender, and ethnicity, and 10,950 for the other predictors in the model. The statistical significance of the Student's t statistic is presented in the fifth column.

Table 33.1 Population averaged model for prevalence of marijuana use

	*b**	*b*	*SE(b)*	*Student's t*	*p*
$R_O{}^2 = .28$, $p = .000$ df = 1672 level 2 df = 10,950 level 1					
Logit(marijuana use)	–	–	–	–	–
Male	.002	.008	.108	.08	.938
Black	−.035	−.221	.153	−1.44	.149
Other	−.023	−.224	.231	−.97	.334
Exposure (lagged)	.248	.137	.009	14.62	.000
Belief (lagged)	−.221	−.130	.010	−12.64	.000
Age	−.026	−.307	.195	−1.57	.116
Age^2	−.189	−2.343	.198	−11.82	.000
Age^3	.148	1.646	.165	9.96	.000
Age^4	−.039	−.440	.161	−2.73	.007
Intercept/Constant	–	−.971	.085	−11.49	.000

Exposure, belief, and a function of age are statistically significant predictors, but gender and ethnicity do not appear to be statistically significant predictors of marijuana use. The strongest predictor of marijuana use appears to be exposure, followed by belief that it is wrong to violate the law. Because age is split into four components, two of which have substantial standardized coefficients, the question arises, what is the contribution of age relative to other variables in the model, particularly exposure and belief, to the explanation of marijuana use? One possible approach to answering this question is to use the technique described in Menard (2004) of multiplying each standardized coefficient by the corresponding zero order correlation to estimate the direct contribution to the explained variance of each predictor, which gives us

$$\Sigma b^{*}r = (-.032)(-.026) + (-.201)(-.189) + (.169)(.148) + (.027)(-.039) = .063.$$

Notice that the last of the four components of the sum is negative (but very small, −.001), not uncommon with correlations and standardized coefficients that are small in magnitude. For exposure, the direct contribution to the explained variance is (.447)(.248) = .111, and for belief it is (−.439)(−.221) = .097. Similarly, the direct contribution of ethnicity is (−.032)(−.035) + (−.027)(−.023) = .002, and for gender it is (.083)(.002) = .0001, a negligible effect.

5 The unit-specific model for prevalence of marijuana use

In contrast to the population averaged model, in the unit-specific model, the concern is with the extent to which a change in an independent variable for a particular observation is associated with a change in the dependent variable for that same case. For example, how much of an impact would an increase in exposure or belief *for a particular individual* have on that individual's marijuana use? This model would typically be more appropriate for (1) developmental change, when different cases in the population had different individual experiences (e.g., different changes in belief) over time, and (2) narrowly based individual level as opposed to population level interventions, when different cases were subjected to different interventions or different levels of intervention. The random effects model includes specific random components for the cases which are not actually estimated, but the parameters (the variance components) of the assumed distribution of the random effects are incorporated into the model (Hardin and Hilbe, 2003, pp. 42–49). As a practical matter, results from estimating population averaged and random effects models are often quite similar, but in some instances there may be interesting differences. Table 33.2, parallel in structure to Table 33.1, presents the results of estimating a random effects model for the prevalence of marijuana use.

A random effects model may include a random intercept β_0, random slopes $\beta_1, \beta_2, \ldots, \beta_K$, or both. In the present analysis, the model was tested separately for random effects in the coefficients for age and the coefficients for exposure and belief, but the random effects for these coefficients were not statistically significant, and thus they are excluded from the model. The variance attributable to the random intercept is statistically significant (p = .000), and the intraclass correlation (the proportion of the total variance attributable to the variance between level 2 units, in this case individual respondents) rho = .518, p = .000. In other words, roughly half of the total variance between *observations* actually occurs between *individuals* (the level 2 units).

For the model in Table 33.2, $R_O^2 = .29$ (p = .000). Once again, exposure, belief, and age are statistically significant, and gender is not statistically significant as a predictor of marijuana use, but this time ethnicity is also a statistically significant predictor. Substantively, the

Table 33.2 Random effects model for prevalence of marijuana use

	*b**	*b*	*SE(b)*	*Student's t*	*p*
$R_O^2 = .29$, $p = .000$ df = 1672 level 2 df = 10,950 level 1 $D_M = 29{,}565.17$ $D_{0(\text{HLM})} = 31{,}691.19$; $D_{0(\text{logit})} = 14.678.19$ $G_M = 2126.02$, df = 9, p = .000 $R_L^2 = .14$ (based on $D_{0(\text{logit})}$ and G_M)					
Logit(marijuana use)	–	–	–	–	–
Male	−.010	−.066	.122	−0.54	.587
Black	−.037	−.366	.160	−2.29	.022
Other	−.024	−.358	.250	−1.43	.152
Exposure (lagged)	.266	.229	.010	23.78	.000
Belief (lagged)	−.229	−.212	.012	−18.19	.000
Age	−.030	−.560	.232	−2.41	.016
Age^2	−.172	−3.348	.249	−13.44	.000
Age^3	.138	2.407	.217	11.08	.000
Age^4	−.035	−.612	.202	−3.04	.000
Intercept/Constant	–	−1.394	.100	−13.90	.000

results are similar to the results from the population averaged model. Based on calculation of the direct contribution of each predictor to the explained variance in the dependent variable, using the same technique described above, the strongest predictors are exposure and belief, followed by age. The effect of ethnicity is statistically significant, with African-Americans in particular being less likely to use marijuana across this part of the life span, but substantively the effect is weak, with ethnicity (both Black and Other) accounting for little more than 0.1% of the variance in marijuana use. The effect of gender is not statistically significant. If we plot the mean predicted values of marijuana use with age, we obtain the result in Figure 33.2. Note that by projecting the plot past age 10, we would obtain a negative predicted probability of marijuana use, illustrating the danger of projecting predicted values using polynomial functions past the range for which the polynomial function was calculated. Since the fourth power of age is negatively signed and dominates the function after about age 30, the same would occur for older ages. Broadly, one can think of the model as (a) describing the pattern of marijuana use across the life cycle from age 12 to age 33, and (b) explaining some of the error in prediction one would obtain using age alone as a predictor in terms of gender, ethnicity, belief, and exposure.

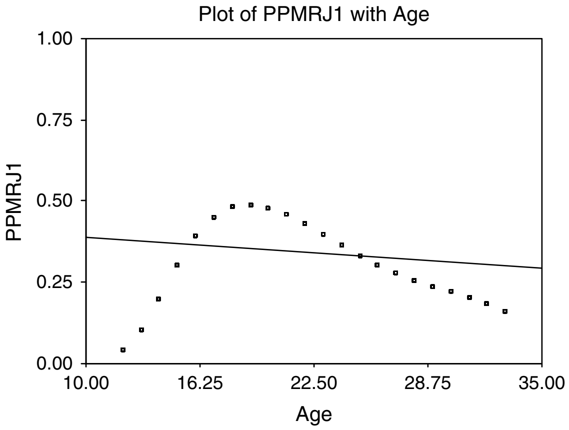


Figure 33.2 Predicted mean prevalence of marijuana use (PPMRJ1) with age

6 Extensions and contrasts

The multilevel models in Tables 33.1 and 33.2 can be extended to include interactions among age and the other level 2 predictors (exposure and belief), interactions among the level 1 predictors (gender and ethnicity), and interactions across levels (e.g., an interaction between exposure and gender), the last being represented by a level 1 effect of gender on the level 2 coefficient for exposure. The extension of the multilevel logistic regression model to polytomous nominal and ordinal variables is relatively straightforward conceptually. A polytomous ordinal dependent variable will typically be modeled with a single set of β parameters using a cumulative logit model with the assumption of parallel slopes, and a polytomous nominal dependent variable with c categories will have c-1 separate functions (equations), as in logistic regression analysis more generally. The polytomous nominal model may also be applicable to ordinal polytomous dependent variables when the parallel slopes assumption is not justified.

Multilevel change modeling for categorical dependent variables is different from two other approaches, panel analysis and event history analysis, in important ways. Panel analysis with logistic regression (see Menard, Chapter 30, in this volume) typically involves fewer than five separate time points, and does not necessarily include a time dimension in the model. For example, in Chapter 30, several of the models in Tables 30.2 and 30.3 included no time dimension at all, and when they did include age as a time dimension, because there were only two waves of data, the age comparisons really involved between-individual rather than within-individual differences in marijuana use associated with age. A time dimension is one of the defining features of event history analysis, and although not absolutely necessary in a multilevel model (age and time may prove to have no statistically significant impacts on the outcome variable), we typically begin in multilevel longitudinal modeling by examining the relationship of the outcome with at least one time dimension. Where multilevel change modeling and (usually) panel analysis differ from event history analysis is in not restricting the analysis to cases defined as being "at risk" of change in the direction of interest. In event history analysis, *initiation* (whether someone who has not previously used marijuana begins to do so, for which only nonusers are at risk) or *suspension* (whether someone who has been using marijuana ceases to do so, for which only users are at risk) would typically be the variables of interest, each with its own separate at risk set of cases. In multilevel analysis and (again usually; see Chapter 30, option d for measuring change in dichotomous dependent variables) panel analysis, it is more typically *prevalence* (simply whether one does or does not use marijuana) that is the dependent variable of interest, and the risk set is typically defined as all respondents. Multilevel analysis, panel analysis, and event history analysis thus present a complementary set of approaches for the analysis of longitudinal categorical data.

7 Conclusion: multilevel logistic regression for longitudinal data analysis

Most of the existing literature on multilevel modeling treats the analysis of discrete dependent variables generally, and logistic regression in particular, as an afterthought or a secondary issue, assuming a focus on interval/ratio dependent variables and partitioning of variance (and by the way, if you are unfortunate enough to have to deal with discrete dependent variables, here is something that may help). Here, the focus has been somewhat different, attempting to strengthen the bridge between multilevel modeling and modeling discrete dependent variables. In this light, greater emphasis has been given than is usual in the multilevel modeling literature to the calculation of appropriate measures for explained variation to assess the overall impact of the predictors on the dependent variable, and standardized coefficients for assessing the relative impact within a model of predictors measured on different scales. The good news is that for the simplest (e.g., random intercept) models, one can relatively easily obtain estimates from existing software packages. The bad news is that (a) in order to know for sure whether the simpler model is adequate, it may be necessary to first examine the more complex models, (b) the statistical routines in general-purpose statistical software packages are not up to the task of estimating the more complex models, and one must rely instead on more specialized multilevel statistical software (an observation also made by Luke, 2004, p. 73), and (c) for more complex models, it may not be possible to obtain maximum likelihood estimates, with all that implies for assessing the relative fit of different models and explained variation based on the criterion actually being maximized in a logistic regression or logit model approach. The principal strength and importance of multilevel logistic regression modeling for longitudinal data is that it takes into account dependencies in the data occasioned by repeated measurement of the same cases (level 2 units) in modeling the relationship of the outcome with time (chronological or age), time-constant attributes, and time-varying covariates.

Software

Analysis of the relationship between marijuana use and the orthogonal polynomial contrast of age was performed using both SAS and SPSS **logistic**. The results in Tables 33.1

and 33.2 were calculated using the statistical package HLM (Raudenbush et al., 2004). Similar results were obtained for the population averaged and random intercept models using Stata **xtlogit**. In presently available software, it may be necessary to calculate the measures of explained variation and standardized logistic regression coefficients used here by hand, using the squared correlation between observed and predicted values of the dependent variable to calculate R_O^2 (Menard, 2000 and 2002), and using the formula described earlier in this chapter (and in more detail in Menard, 2002 and 2004) for standardized coefficients. Population averaged models may be calculated using software other than dedicated multilevel software when that software permits adjustment for dependency in repeated measures within cases. For ordinal polytomous dependent variables, a test of the parallel slopes assumption may be available in the particular multilevel modeling software being used; if not, the best option for testing the assumption of parallel slopes may be to calculate the model using standard statistical software such as SAS, SPSS, or Stata, solely for the purpose of choosing between the alternative polytomous multilevel models, then calculating the selected model using the appropriate multilevel software.

References

Bijleveld, C. C. J. H. and van der Kamp, L. J. T., with Mooijaart, A., van der Kloot, W. A., van der Leeden, R. and van der Burg, E. (1998). *Longitudinal Data Analysis: Designs, Models, and Methods.* London: Sage.

Elliott, D. S., Huizinga, D. and Menard, S. (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems.* New York: Springer-Verlag.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized Estimating Equations.* Boca Raton, FL: Chapman & Hall/CRC.

Luke, D. A. (2004). *Multilevel Modeling.* Thousand Oaks, CA: Sage.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *American Statistician*, 54: 17–24.

Menard, S. (2002). *Applied Logistic Regression Analysis,* second edition. Thousand Oaks: Sage.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *American Statistician*, 58: 218–223.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods,* 2nd edn. Thousand Oaks, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F. and Congdon, R. (2004). *HLM6: Hierarchical Linear and Nonlinear Modeling.* Lincolnwood, IL: Scientific Software International.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* London, UK: Sage.

Part VII

Time Series Analysis and Deterministic Dynamic Models

This page intentionally left blank

Chapter 34

A brief introduction to time series analysis

Scott Menard

1 Introduction

This chapter provides a brief introduction to time series analysis, focusing on autoregressive integrated moving average (ARIMA) analysis of time series. A *time series* is a set of repeated measurements of the same variable taken on the same unit of analysis (e.g., an individual, city, nation; more generally, a *subject* or a *case*) for two or more points in time. As used in the social and behavioral sciences, *time series analysis* typically refers to a large number of observations taken on a single case, typically at equal measurement intervals, possibly on more than one variable. Strictly speaking, even if we have a large number of cases N and a smaller number of time periods T we are, in fact, performing time series analysis, but the terminology used is that of two-wave or multiwave panel analysis, latent or multilevel growth curve or change analysis, and event history analysis, topics covered in other chapters in this volume. Here we focus on analysis of a single case ($N = 1$) for a large number of time periods (typically $T > 20$ and preferably $t > 50$ for most purposes, sometimes $T > 100$).

Time series analysis will always have at least one of three goals: description, explanation, or forecasting. In principle, one can perform time series analysis on variables at any level of measurement (dichotomous, nominal, ordinal, interval, or ratio-scaled variables), but in practice the techniques that are commonly described as time series analysis are applied to more or less continuous quantitative (interval or ratio-scaled) outcomes, with equal intervals between successive measurements. Focusing, then, on quantitative outcomes, we may use time series analysis to *describe* the level or value of the outcome variable as a function of (1) time itself, (2) past levels or values of the same outcome variable, (3) past and present values of a random change, or a *random shock,* to the level or value of the outcome, or (4) some combination of all three. In addition, we may attempt to *explain* the value of the outcome in terms of (5) one-time nonrandom shocks to the series, as in an intervention analysis in which an intervention or a policy change is tested to see whether it has an impact on the time series, or in terms of (6) one or more *time-varying covariates,* which are themselves time series and can be represented as inputs for the outcome variable. Finally, once we model the effects of time, past values of the outcome, random shock, and covariates, we can use that model to *forecast* future values of that outcome.

An important consideration in time series analysis is whether the time series is *stationary,* an issue to which we shall return repeatedly in this chapter. As described in any number of standard texts on time series analysis, a stationary time series has a fixed mean and a constant variance about that mean. If, in addition, the *autocovariance structure* (the covariance between z_t and z_{t-s} for $s = 1, 2, \ldots, S < T$) is also constant, the time series is said to be *strictly stationary*. *Nonstationary* time series exhibit either changing variance or changing mean, or both. A changing mean, or *trend,* may involve deterministic changes that are a function of time, predictable from past values of the time series; or they may be stochastic changes, the cumulative result of a series of random shocks which, at random, have more often "pushed" the time series up (or down) rather than in the opposite direction. Particularly in the case of bivariate time series analysis, if both series have similar stochastic trends, the stochastic trends may produce a spurious positive correlation inflating estimates of how well one series predicts or explains the other, or it may attenuate a true negative correlation between the two series. Alternatively, if the stochastic trends are in opposite directions, they may either attenuate a true positive correlation or produce a spurious negative correlation between the two series.

1.1 Notation

The notation used in describing equations for time series analysis is not standard across different sources. In this chapter, the following notation will be used; capital letters refer to variables, lower case letters to values of variables or constants.

T is the number of times, or equivalently, the number of observations in the time series; $t = 1, 2, \ldots, T$ may also be a variable in the equation for the dependent variable.

Z is the dependent variable; z_t is the dependent variable measured at time $t = 1, 2, \ldots, T$.

$\{Y_{k,t}\}$ is a set of K time-varying covariates; $y_{k,t}$ is the time-varying covariate Y_k, $k = 1, 2, \ldots, K$ measured at time $t = 1, 2, \ldots, T$.

$\{X_{j,t}\}$ is a second set of J covariates, possibly but not necessarily time-varying; $x_{j,t}$ is the time-varying covariate X_j, $j = 1, 2, \ldots, J$ measured at time $t = 1, 2, \ldots, T$.

A residual at time $t = 1, 2, \ldots, T$ is denoted e_t.

When the residual e_t is a white noise residual (random shock) it is denoted a_t.

Coefficients for lagged endogenous variables $z_{t-1}, z_{t-2}, \ldots, z_{t-p}$ are denoted ϕ_p, $p = 1, 2, \ldots, \pi < T$.

Coefficients for past random shocks $a_{t-1}, a_{t-2}, \ldots, a_{t-q}$ are denoted θ_q, $q = 1, 2, \ldots, \xi < T$.

Coefficients for time-varying covariates are denoted $\beta_{k,t}$ or $\gamma_{j,t}$ where k or j refers to the covariate Y_k or X_k, $k = 1, 2, \ldots, K, j = 1, 2, \ldots, J$, and $t = 1, 2, \ldots, T$ is the time at which the covariate was measured.

Coefficients for time (t) are denoted λ, subscripted for different functions or transformations of time, f(t); in general, $z_t = \lambda_0 + \lambda_1 f_1(t) + \lambda_2 f_2(t) + \cdots + \lambda_M f_M(t)$ for $m = 1, 2, \ldots, M$ different functions of time.

2 Describing or modeling the outcome as a function of time

Some outcomes can be modeled as direct functions of time, with time conceived as a predictor but not a cause of the outcome. The compounding of interest for a savings bond is one example. A fixed amount is deposited, and the account grows at a fixed rate as long as the bond is held. The direct "causes" of this growth lie in (a) the agreement upon the terms of the contract between the buyer and the seller of the bond and (b) the decision of the bondholder to continue to hold the bond, but the growth in the account itself is a deterministic function of time, and the only information needed to predict the total value of the account at any given time is (1) the value of the account at some previous time, (2) the amount of time lapsed since the time for which the previous value was obtained, and either (3a) the interest-rate terms for the account or (3b) a sufficient number of values of the account at different prior times to calculate the interest rate, given general knowledge of the interest-rate terms of the account (e.g., annual, daily, or other compounding of interest), even without knowing the specific interest rate. With (3a), knowledge of the interest-rate terms, we can directly calculate the value of the account at a given time. With (3b), knowledge of the time series of prior values of the account (plus a knowledge of the mathematics of compound interest), we can select a function of the appropriate form but with the interest rate as an unknown variable, use the time series data to calculate the interest rate, and from that information we can calculate the value of the account at any given time.

To illustrate, assume that the account earns some unknown rate of interest, x, per year (or whatever time interval is appropriate). Then given an initial value of the account at time zero, z_0, and measurements for at least one other time, t, where the measurement interval between times $t = 1, 2, \ldots, T$ corresponds to the interval at which interest is compounded, we know that $z_t = z_0(1+x)^t$. In the case of a purely deterministic model for which the function is known, we need only one time point to solve for x: $x = (z_t/z_0)^{1/t} - 1$, and given x we can calculate what z_t will be for any time after (or for that matter, all possible values of z_t prior to) $t = 0$. With the information about the form of the curve and values of z_t at different times, we have fit the curve for the outcome (the value of the account) by calculating the unknown parameter x.

More generally, the precise function defining the curve (or straight line) to be fit will not be known, and the first step in the analysis will be to estimate what that function might be. For this we need values of the outcome for many more than two points in time. Assuming no systematic or random error, a longer time series should indicate whether the values of the outcome increase, decrease, or oscillate over time, whether (and if so how many times) the curve changes direction from increasing to decreasing or *vice-versa*, and based on this information we may be able to choose a suitable function for the curve, but for any given curve, there will be an indefinitely large number of possibilities.

One possible basis for choosing one particular curve may be to have some idea of the underlying process, as in the previous example of compound interest. If we have no real knowledge of the underlying process, however, we may use the knowledge that any curve can be approximated by a polynomial of order $M < T$, and estimate the parameters for the equation. For a model with an additive effect of time (the value is changed by adding a fixed function of time),

$$z_t = \lambda_0 + \lambda_1(t) + \lambda_2(t^2) + \cdots + \lambda_M(t^M),$$
$$\text{where } \lambda_0 = z_0 \tag{1}$$

that is, z_t is equal to the intercept when $t = 0$.

Alternatively, the amount may be increased by some multiplicative function of time. There are several different multiplicative functions that could be used, only some of which are linear in their parameters (and hence can be estimated using ordinary least squares linear regression). One possibility is that the relationship can be expressed as an exponential function of a polynomial function of time,

$$z_t = \exp(\lambda_0 + \lambda_1 t + \lambda_2 t^2 + \cdots + \lambda_M t^M) \tag{2}$$

in which $\exp(x) = e^x$ where e is the base of the natural logarithm, and which translates into

$$\ln(z_t) = \lambda_0 + \lambda_1 t + \lambda_2 t^2 + \cdots + \lambda_M t^M \tag{3}$$

and here $\ln(z_0) = \lambda_0$, and $\ln(z_t)$ is expressed as a polynomial function of time that is linear in the parameters. Other possibilities exist involving other transformations of z_t and other functions of time, and these possibilities can be modeled using existing time series software (e.g., SPSS CURVEFIT).

The order of the polynomial $m = 1, 2, \ldots, M$ may be guessed from visual inspection of the curve for the number of changes in the curve between positive and negative trends, and M should be set to at least one more than the number of changes in trend. Alternatively, one can test polynomial functions of different orders to see which of the coefficients $\lambda_1, \lambda_2, \ldots, \lambda_M$ can be dropped. For this purpose, because the various powers of t if untransformed will be highly collinear, it will be necessary to use an *orthogonal contrast* (e.g., Scheffé, 1959) of the appropriate order for t in order to ascertain which of the $\lambda_1, \lambda_2, \ldots, \lambda_M$ are statistically significant. There are different possible criteria for deciding which of the λ coefficients to retain, including (a) a hierarchical approach, retaining all coefficients, up to and including the last statistically significant coefficient, regardless of the statistical significance of coefficients between the first and last statistically significant coefficients, and (b) a nonhierarchical approach, dropping all coefficients that are not statistically significant but retaining all coefficients that are statistically significant, either in a single-step procedure (a single assessment of statistical significance, controlling for all of the other coefficients) or in a stepwise process that eliminates the least statistically significant coefficient first, then re-estimates the model and repeats the process until all of the remaining coefficients meet some criterion of statistical significance. One can also choose (c) to limit the powers of t *a priori* to some maximum, for example 4 or 6. This last approach stems from the observation that in the natural sciences, phenomena involving exponents larger than 3 or 4 are exceedingly rare, plus the concern that, by including higher powers of t, we may be overfitting the model, fitting random error variation rather than true variation in the outcome. Alternatively, based on the assumption that the polynomial function is only *approximating* the true function, we may choose to include the higher powers of t if they appear to be statistically significant.

This approach may work well in the absence or near absence of random error, but any substantial amount of random error (or *noise*) in the time series may produce apparent increases or decreases at different points on the curve that lead to overestimation or underestimation of the value of the outcome, and more apparent changes in trend than are really characteristic of the true relationship between the outcome and time. To address this problem and the associated risk of overfitting the curve, *curve smoothing* techniques may be applied to reduce the random error at each point on the curve. One of the more common techniques of smoothing is to represent each point on the curve by a *moving average*. For example, z_t may be represented as the average $(z_{t-1} + z_t + z_{t+1})/3$. Alternatively, one could take a longer moving average (e.g., 5 instead of 3 time points). The use of a moving average relies on the expectation that the sum of a series of random errors will be equal to zero, and hence the random errors from previous and subsequent values of z_t will cancel out the random error in z_t. One could also weight the moving average, for example representing z_t by $(z_{t-1} + 2z_t + z_{t+1})$, giving more weight to the current value of z_t. This arises naturally when using *double moving averages* for smoothing. A double moving average is simply a moving average of moving averages, where one first takes a moving average of the z_t producing new values z_t', then takes the moving average of the z_t' in turn. As illustrated in Yaffee and McGee (2000, p. 22), using a double moving average of length 3 for both the first and the second moving average, z_t is expressed as $z_t'' = (z_{t-2} + 2z_{t-1} + 3z_t + 2z_{t+1} + z_{t+2})/9$. Once the curve has been smoothed using single or double moving averages, it may be possible to identify the appropriate function or to calculate an appropriate polynomial approximation to the function to describe the series. Once this is done, one can calculate residuals $e_t = z_t - \hat{z}_t$ where $\hat{z}_t$ is the estimated value of z_t based on the fitted curve.

These residuals can then be subjected to further analysis.

An example of curve fitting in the social sciences is the ambitious attempt by Hamblin et al. (1973) to describe processes of social change. Hamblin et al. model time series for such dependent variables as industrial startups, adoption of innovations such as hybrid corn and new drugs, general strikes, political assassinations, air passenger miles, gasoline consumption, and gross national product. Using a combination of differential equations and curve fitting, they derive a series of models, including linear, exponential, and logistic models, to describe (and provide theoretical explanations for) the evolution of these variables over time. The focus on these models is (a) assumptions about the nature of the process driving change in the dependent variable over time, and (b) given the assumption of constancy in that process, modeling the dependent variable as a function of time, with allowances for change in the process which may result in changes in the direction or other characteristics of the curve for different "epochs" or periods within which the assumptions about the process are valid. For Hamblin et al., the theory generates expectations regarding the relationship of the dependent variable with time, which are then tested by fitting the appropriate function of time to the dependent variable.

3 Describing or modeling the outcome as a function of present and past random shocks

It is possible that there is no deterministic trend in the model, or that any such trend has been accounted for and removed from the model (e.g., by fitting the curve and calculating the residuals, as described above, leaving the residuals for further analysis). There may still remain systematic patterns to be described in the series. One possibility is that the outcome, z_t, can be described as the function of a random shock at time t, a_t, plus the lingering effects of a series of some number q of past random shocks a_{t-1}, $a_{t-2}, \ldots, a_{t-q}$, occurring to the series after some initial time t_0 but prior to t (with z_t assumed to be measured immediately after a_t occurs). The *moving average process of order q, MA(q),* as described in Box and Jenkins (1970), is modeled as $z_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q-1}$, a linear function of the current random shock and the past q random shocks. The θ coefficients are parameters to be estimated, and q is to be identified based on the *autocorrelation function (ACF)* and *partial autocorrelation function (PACF).* The θ coefficients represent the magnitude of each of the past random shocks, and q represents the length of time after which the effect of a previous random shock is completely (or at least for all practical purposes) dissipated. It is assumed that the random shocks have a mean of zero, i.e., $E(a_t) = 0$; if, instead, they vary around a nonzero mean, the series is said to exhibit *drift*, random variation about a nonzero mean, and the series is nonstationary.

In modeling an MA(q) process, the first order of business is to identify q using the ACF and PACF. The ACF represents the correlation of each value of Z with itself at different *lags:* z_t with z_{t-1} for all t (lag 1); z_t with z_{t-2} for all t (lag 2); and so forth, for some number of lags τ, typically with $\tau < .5T$, and for longer series, τ may be substantially less than .5T (e.g., 40 or fewer lags). The ACF is calculated as

$$r_\tau = [1/(T-\tau)] \sum_{t=1}^{T-\tau} (z_t - \bar{z})(z_{t-k} - \bar{z}) / [1/(T-\tau)] \times \sum_{t=1}^{T} (z_t - \bar{z})^2 \qquad (4)$$

where $\bar{z}$ is the mean of the time series and τ is the number of lags for which autocorrelations are calculated. The standard error for r_τ is

$$SE(r_\tau) = \sqrt{(1/N)[1 + 2 \sum_{L=0}^{\tau-1} r_L^2]} \qquad (5)$$

where L is the lag and ${r_0}^2$ is defined as ${r_0}^2 = 0$. The PACF is the partial autocorrelation of y with itself at each lag, controlling for autocorrelations at intervening lags: z_t with z_{t-2} controlling for the correlation of z_t and z_{t-2} with z_{t-1}; z_t with z_{t-3} controlling for the correlations of z_t and z_{t-3} with z_{t-1} and z_{t-2}; etc. The PACF is calculated in an iterative process, beginning with $\mathrm{PACF(lag\ 1)} = r_1$, $\mathrm{PACF(lag\ 2)} = (r_2 - r_1^2)/(1 - r_1^2)$, and continuing for higher order partial autocorrelations expressed in terms of the current autocorrelation and lower order autocorrelations. The standard error for the partial autocorrelation is defined as $\mathrm{SE(PACF)} = 1/\sqrt{T}$.

The characteristic pattern for an MA(q) process is that (1) the ACF "spikes" (is large enough to be statistically significant, and is markedly larger than other autocorrelations) at the first q lags, where q may in principle be 1, 2, . . . , T-2, and beyond lag 1 the ACF is otherwise small and not statistically significant; and (2) the PACF declines fairly quickly as the length of the lag increases. Once q has been identified, we proceed with the estimation of the θ coefficients. In an MA(q) model these coefficients must satisfy the condition of *invertibility*. For an MA(1) process, invertibility exists when the absolute value of θ_1 is less than 1: $|\theta_1| < 1$, that is, $-1 < \theta_1 < 1$. For an MA(2) process, there are three conditions: (1) $\theta_1 + \theta_2 < 1$; (2) $|\theta_2| < 1$; and (3) $\theta_2 - \theta_1 < 1$. More complicated conditions for invertibility exist for higher order MA processes, but MA processes with $q > 2$ are relatively rare in practice.

4 Describing or modeling the outcome as a function of past values of the outcome (plus a_t)

Even when there is no deterministic trend in the data, a stochastic trend may appear as a result of the cumulative impact of random shocks in the series, even if there are no lingering effects of the random shocks. The simplest model for this stochastic trend is the *random walk,* in which a random shock added to the most recent value of the series produces the current value of the series: $z_t = z_{t-1} + a_t$ or equivalently $(z_t - z_{t-1}) = a_t$. This process is also described as an *integrated process of order d = 1, or I(1),* indicating that the difference between two adjacent values of the outcome is a random shock, a_t. Higher order integrated processes are also possible, starting with an integrated process of order 2, I(2), for which $(z_t - z_{t-1}) - (z_{t-1} - z_{t-2}) = z_t - 2z_{t-1} + z_{t-3} = a_{t.}$ In practice, integrated processes of order higher than $d = 2$ are rare, but although calculation of such processes becomes increasingly tedious, they are possible. As is evident from the foregoing, in the integrated process, the value of the outcome depends directly on its past value(s) plus a random shock. The symbol ∇^d is often used as a shorthand for a difference corresponding to an integrated process of order d, such that $\nabla^1 = (z_t - z_{t-1})$, $\nabla^2 = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2})$, etc. For an integrated process of order 1, I(1), probably the most common integrated process encountered in practice, the characteristic pattern of the ACF is that it declines gradually with time, while the PACF spikes for lag 1. Integration or differencing is limitedly parallel to curve fitting using a polynomial function; taking the first difference removes the linear trend, taking the second difference removes the quadratic trend, and so forth, but unlike polynomial (or other) curve fitting, the use of first or higher order differences does not inform us about the pattern or shape of the observed relationship of the outcome variable with time.

As an alternative to the integrated model using first or higher order differencing, one or more past values of the outcome may be hypothesized to have a lingering effect on z_t that is not adequately captured by the simple integrated process; instead we model an *autoregressive process of order p, AR(p),* where p is the number of past values of the outcome on which the current value, z_t, depends. In an AR(1) process, $z_t = \phi_1 z_{t-1} + a_t$. This looks a lot like the random walk, except for the coefficient, ϕ_1, which adds a coefficient to be estimated to represent the dependence of z_t on z_{t-1}. If $\phi_1 = 1$, we have not an AR(1) process but instead an I(1) process; for an AR(1) process, it must be the case that $-1 < \phi_1 < 1$. For an AR(2) process, $z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t$. Like the MA(q) process, the AR(p) process may be identified based on the ACF and PACF, with the characteristic pattern of an AR(p) process being somewhat similar to the pattern for an I(d) process and a mirror image of the pattern for an MA(q) process. For an AR(p) process, it is the ACF rather than the PACF that declines fairly quickly with time, and the PACF rather than the ACF that has “spikes” at the p lags for which we can expect the effects of prior values of the outcome to be significant. The difference between this and the I(d) process is that for the I(d) process, the ACF declines more gradually with time than for the AR(p) process.

Also parallel to the MA(q) process, there are conditions that must be met for the AR(p) process to be estimated, but here they are conditions of *stationarity* rather than invertibility. For an AR(1) process, stationarity requires that $|\phi_1| < 1$ [and as noted above, if $\phi_1 = 1$, the supposed AR(1) process is actually an I(1) process]. Note also that, as implied above, the autocorrelation may be either positive or negative; i.e., for some time series, there is a tendency for oscillation in the value of z_t such that a high value of z_t tends to be followed by a low value of z_t, characteristic of a process that strives to reach an equilibrium. For an AR(2) model, stationarity requires that (1) $\phi_1 + \phi_2 < 1$; (2) $|\phi_2| < 1$; and (3) $\phi_2 - \phi_1 < 1$. More complicated AR(p) models have more complicated conditions of stationarity, but in practice AR(p) models of order greater than $p = 2$ are relatively rare.

4.1 The Dickey-Fuller and augmented Dickey-Fuller tests

As noted earlier, stationarity in a time series means that there is a constant mean, and constant variance about that mean. In particular, in order for a time series to be stationary, it is necessary to account for and remove any trend, stochastic or deterministic. A deterministic trend may be removed by curve fitting using the methods described in Section 1 of this chapter, calculating the residuals, and performing the time series analysis on the residuals of the fitted curve. Stochastic trends, as in a random walk, may often be removed by taking first or second differences. It is possible to test for the presence of an I(1) component to the time series using the Dickey-Fuller or augmented Dickey-Fuller tests. The Dickey-Fuller tests calculate an autoregressive model and test whether the coefficient ϕ_1 is statistically different from one. If it is not, it will be necessary to difference the series to achieve stationarity. Different versions of the augmented Dickey-Fuller tests provide evidence regarding the evidence of random walk without drift, random walk with drift, or trend in a time series. The Dickey-Fuller or augmented Dickey-Fuller tests and other tests for whether $\phi_1 = 1$ (or, as it is sometimes described, whether the series has a unit root) are commonly available in time series analysis software.

5 The ARIMA(p,d,q) model

A time series may be described as a combination of autoregressive, integrated, and moving average components: AR(p) plus I(d) plus MA(q), or ARIMA(p,d,q). Note that the AR(p) and I(d) are both processes that model the current value of the outcome as a function of one or more past values of the outcome. One might thus expect to find one or the other, but not both, processes operating in a single model, but, in principle, all three processes could be combined in a single model. For example, an ARIMA(1,1,1) model may be written $(z_t - z_{t-1}) = \phi_1(z_{t-1} - z_{t-2}) + \theta_1 a_{t-1} + a_t$, or $\nabla^1 z_t = \phi_1 \nabla^1 z_{t-1} + \theta_1 a_{t-1} + a_t$. In English, the *difference* between the current and the immediately past value of the outcome is equal to the previous difference (the difference between the immediately past value of the outcome and the value of the outcome immediately prior to that) multiplied by ϕ_1, plus the value of the immediately past random shock multiplied by θ_1, plus the current random shock a_t. Stated another way, the difference between the present and the immediately past value of z_t is equal to (1) a random shock a_t; plus (2) a linear function of the previous random shock, $\theta_1 a_{t-1}$, the MA(1) component; plus (3) the previous value of the outcome, the I(1) component; plus (4) a linear function of the previous difference in the values of the outcome, $\phi_1(z_{t-1} - z_{t-2})$, the AR(1) component.

As noted earlier, the order of each of the three types of processes is relatively small in practice. For AR(p) and MA(q) processes in particular, there is a theoretical reason why this should be the case. Briefly, an AR(p) model converges to an MA(1) model as p becomes infinitely large, and an MA(q) process converges to an AR(1) model as q becomes infinitely large. The larger the order of p or q, the better the AR(p) or MA(q) processes can be represented by a lower order process of the opposite type. Identification of which one or more of the processes best describes the time series is based on examination of the ACF and PACF, as described earlier. It is not always the case, however, that the ACF and PACF unambiguously identify p, d, and q for a time series. In this case, it may be useful to compare different possible models for how well they describe the time series. The Dickey-Fuller test is one criterion to consider. A second is whether all of the coefficients (ϕ and θ) in the model are statistically significant; if not, the model is rejected, and a different model, without the nonsignificant coefficients, is more appropriate. There are also several general tests of model fit, including the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Schwartz Bayes criterion (SBC), available when the parameters of the model are estimated using maximum likelihood (see, e.g., Wei, 2006, pp. 156–157). The lower the AIC, BIC, or SBC, the better the fit of the model. As noted in Wei (2006), the BIC may be preferable to the older AIC, because the AIC has a tendency to overparameterize the model. As noted in Tabachnick and Fidell (2007, 18.16), the AIC has the property that the difference in two AIC statistics is distributed as a χ^2 statistic with appropriate degrees of freedom, when one model is nested within the other, thus giving us a statistical test of the difference in the AIC for the two models: $\chi^2 = \text{Larger AIC} - \text{Smaller AIC}$. For further details on these and other tests used in time series analysis, see, e.g., Cromwell, Labys and Terraza (1994) and Cromwell, Hannan, Labys and Terraza (1994).

6 Example: IBM stock prices

Figure 34.1 is a chart of IBM common stock closing process taken daily from May 17, 1961 to November 2, 1962, Box and Jenkins' (1976, p. 526), Series B, a well-known example in time series analysis. The series appears to increase, then decrease, then increase again (and perhaps decrease thereafter), and seems like a reasonable candidate for polynomial curve fitting. Here, we calculate a fourth order polynomial to account for the three changes in direction

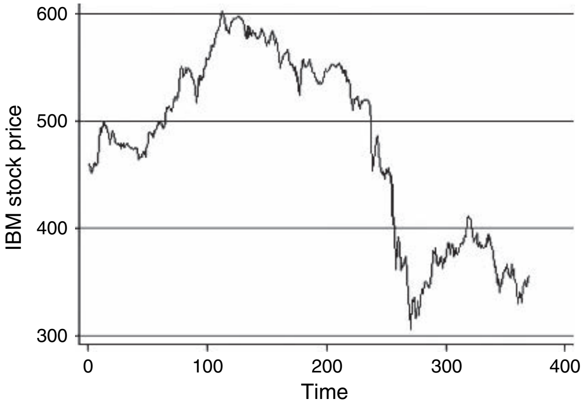
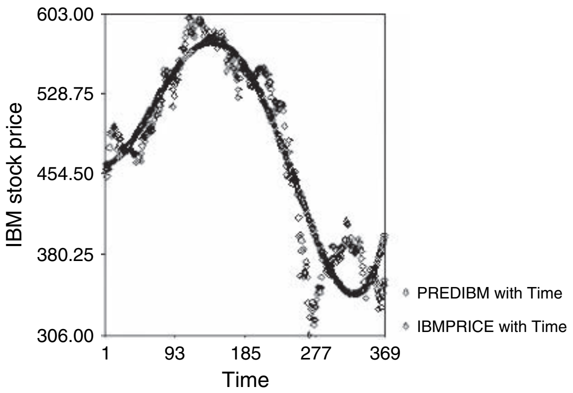


Figure 34.1 IBM common stock closing prices, daily, May 17, 1961 to November 2, 1962
Source: Calculated from Box and Jenkins (1976, p. 526) series B



Equation: Price = 514.488 − 239.633t − 689.325t^2 + 391.874t^3 + 170.080t^4
R^2 = .891, P = .000; all coefficients statistically
significant (p = .000).

Figure 34.2 Quadratic model for IBM stock price

in the curve. The values of the time dimension have been transformed to reduce the collinearity among the different powers of time in the polynomial equation.

Figure 34.2 presents the results of fitting the fourth order polynomial to the IBM stock price data. The darker curve represents the predictions generated by the equation: Price = 514.488 − 239.633t − 689.325t^2 + 391.847t^3 + 170.080t^4. Again, the variable t (time) has been transformed to reduce collinearity, so the numeric values of the coefficients are not themselves meaningful; what they do is (a) indicate that the coefficients for the linear and quadratic terms are negative, while the coefficients for the cubic and quartic terms are positive, and (b) produce the predicted curve in Figure 34.2. This predicted curve is overlaid with the observed values of the curve from Figure 34.1. We can see that prediction is good earlier in the series, but the fourth order polynomial fails to capture the pattern later in the series, suggesting that a higher order polynomial may be necessary to better approximate the series. The polynomial curve provides a description of the *pattern* of the IBM stock prices over time; but absent some theory of why we would expect the process to follow the pattern represented or approximated by a fourth or higher order polynomial, it offers little insight into the *process* by which the pattern was generated.

Figure 34.3 presents the autocorrelation and partial autocorrelation functions for the same IBM price data. The autocorrelation function in Figure 34.3 appears to display the very gradual decline characteristic of an integrated process, and the partial autocorrelation function spikes at lag 1, suggesting an I(1) process. An augmented Dickey-Fuller test confirms that the series appears to have a unit root. In Table 34.1, the ARIMA(0,1,0) model is calculated, along with three alternative models: an ARIMA(1,0,0) model or equivalently an AR(1) model, to further illustrate why the I(1) model is preferable to the AR(1) model here; an ARIMA(0,1,1) model representing a random walk with drift; and an ARIMA(1,1,0) model, which will be of more interest shortly. Results of some other models are summarized below Table 34.1 to explore the possibility that a higher order autoregressive or moving average parameter might improve the model. AIC and BIC statistics were included in the output and are also presented here. As noted earlier, the AIC tends to overparameterize ARIMA models, but it is included here at least for illustrative purposes because of its widespread use (in both time series analysis and other applications).

First, note the magnitude of the ϕ coefficient in the ARIMA(1,0,0) model. As indicated earlier, the augmented Dickey-Fuller test indicates that this coefficient is not statistically significantly different from one, indicating that an I(1) model is more appropriate than an AR(1) model. Second, comparing the AIC and BIC across the models, the ARIMA(0,1,0) model has the lowest BIC but not the lowest AIC; instead, the AIC is lowest for the ARIMA(0,1,1) model, and the AIC for the ARIMA(1,1,0) model is practically identical to that for the ARIMA(0,1,1) model. Which model is best? According to the BIC, which is lowest for the ARIMA(0,1,0) model, the random walk without drift is not only the most parsimonious

```
. corrgram ibmprice

                                            -1       0       1 -1       0       1
 LAG       AC       PAC        Q     Prob > Q [Autocorrelation]  [Partial Autocor]
---------------------------------------------------------------------------------
1        0.9934   0.9991     367.11  0.0000          |-------           |-------
2        0.9859  -0.0871     729.69  0.0000          |-------           |
3        0.9781   0.0071    1087.5   0.0000          |-------           |
4        0.9710   0.0533    1441.1   0.0000          |-------           |
5        0.9638   0.0247    1790.5   0.0000          |-------           |
6        0.9562   0.0183    2135.3   0.0000          |-------           |
7        0.9480  -0.1270    2475.2   0.0000          |-------          -|
8        0.9395  -0.0477    2809.9   0.0000          |-------           |
9        0.9303  -0.0248    3139     0.0000          |-------           |
10       0.9222   0.0607    3463.2   0.0000          |-------           |
11       0.9141  -0.0504    3782.8   0.0000          |-------           |
12       0.9058  -0.0943    4097.4   0.0000          |-------           |
13       0.8974  -0.0283    4407.1   0.0000          |-------           |
14       0.8892   0.0733    4712     0.0000          |-------           |
15       0.8809  -0.0816    5012.1   0.0000          |-------           |
16       0.8724   0.0566    5307.2   0.0000          |------            |
17       0.8632  -0.1464    5597     0.0000          |------           -|
18       0.8530  -0.1037    5880.8   0.0000          |------            |
19       0.8428  -0.0057    6158.6   0.0000          |------            |
20       0.8320  -0.0796    6430.2   0.0000          |------            |
21       0.8207  -0.0861    6695.2   0.0000          |------            |
22       0.8099   0.0730    6953.9   0.0000          |------            |
23       0.7988   0.0445    7206.4   0.0000          |------            |
24       0.7873  -0.0674    7452.4   0.0000          |------            |
25       0.7751   0.0096    7691.5   0.0000          |------            |
26       0.7634  -0.0125    7924.1   0.0000          |------            |
27       0.7523   0.1546    8150.6   0.0000          |------            |-
28       0.7418   0.0427    8371.5   0.0000          |-----             |
29       0.7316  -0.0614    8587     0.0000          |-----             |
30       0.7209  -0.0365    8796.9   0.0000          |-----             |
31       0.7100  -0.0683    9001.1   0.0000          |-----             |
32       0.6988  -0.0385    9199.5   0.0000          |-----             |
33       0.6883   0.0434    9392.5   0.0000          |-----             |
34       0.6787   0.0330    9580.7   0.0000          |-----             |
35       0.6692   0.0256    9764.3   0.0000          |-----             |
36       0.6592  -0.0027    9942.9   0.0000          |-----             |
37       0.6486   0.0183   10116     0.0000          |-----             |
38       0.6376  -0.0477   10285     0.0000          |-----             |
39       0.6271   0.0420   10448     0.0000          |-----             |
40       0.6164  -0.0703   10606     0.0000          |----              |
```

Figure 34.3 Correlogram for IBM daily stock prices

but also the best fitting model. The statistically significant coefficients in the ARIMA(0,1,1) and ARIMA(1,1,0) models, however, suggest that the series contains something more than a random walk without drift. The AIC and BIC are practically identical for the ARIMA(0,1,1) and ARIMA(1,1,0) models, but the tiny difference that does exist favors the ARIMA(0,1,1) random walk with drift model. This is the conventional conclusion (Box and Jenkins, 1970, p. 186, Table 6.4) regarding the IBM stock price series. An important point here is that identification of the appropriate parameterization for a time series model is not always entirely clear, and different criteria may lead us to select different models.

As noted earlier, fitting a fourth order polynomial to the IBM price data provides a reasonable description of the pattern, but little insight

Table 34.1 Comparison of fit statistics for IBM stock price models

Dependent variable	*Statistics/ Coefficients*	*ARIMA (0,1,0)*	*ARIMA (1,0,0)*	*ARIMA (0,1,1)*	*ARIMA (1,1,0)*
IBM stock price	AIC	2506.188	2519.405	2505.483	2505.485
	BIC	2514.004	2531.137	2517.207	2517.210
	DF	2	3	3	3
	ϕ_1 (p)	–	.996 (.000)*	–	.085 (.015)*
	θ_1 (p)	–	–	.085 (.015)*	–
	constant	−.279 (.474)	438.2 (.000)	−.279 (.531)	.280 (.537)

ARIMA(2,0,0): $\phi_1 + \phi_2 = 1$
ARIMA(0,0,2): $\theta_1 + \theta_2 > 1$
ARIMA(0,2,0): AIC = 2721.572, BIC = 2729.283, 2df
ARIMA(0,0,1): AIC = 3863.924, BIC = 3874.657, 3df; $\theta_1 = .931$, p = .000

into the process. By contrast, the ARIMA model does the opposite. Based on this analysis (just a replication of a classic example), the ARIMA model indicates that the process generating the series is simply a nondeterministic function of random shocks to the series, with some lingering effect of the immediately previous random shock. This is the definition of a random walk with drift. However, knowing that the process is a random walk by itself tells us nothing about the pattern of the stock price over time, and might even lead to the mistaken conclusion that there was no trend in the stock price. Instead we have a nonlinear stochastic trend, better *described* (as opposed to *explained*) by the polynomial equation than by the specification of the model as ARIMA(0,1,0). The curve fitting and ARIMA approaches thus provide us with different and complementary information about the pattern and the process, respectively, of the time series.

7 Example: homicides in three midwestern states

In Figures 34.4, 34.5, and 34.6, ACFs and PACFs are presented for the homicide rate per 100,000 people from 1933 to 1980 in Ohio, Wisconsin, and Illinois, using data from Kohfeld and Decker (1990). The mean has been subtracted from each of the time series to center the series. The patterns in the three series are superficially similar: the ACF declines more or less rapidly, and the PACF has a spike at lag 1 for all three series. This appears to be the only spike for the Ohio data, while the data for Illinois suggest a possible second spike, and the data for the PACF for Wisconsin trail off more slowly. Underlying these patterns are three different models for the process generating the three series.

Table 34.2 presents the same ARIMA models for homicides in Ohio, Wisconsin, and Illinois as were presented for the IBM stock price series. Once again, the ARIMA(1,0,0) model is presented to illustrate the need for an I(1) component to the model. In all three ARIMA(1,0,0) models, the ϕ coefficient is large, and, based on the augmented Dickey-Fuller tests, not significantly different from one. For Ohio in Table 34.1, the ARIMA(0,1,0) model is clearly the best choice of the models presented. In both the ARIMA(0,1,1) and the ARIMA(1,1,0) models the coefficients (θ and ϕ) do not attain statistical significance, and although the AIC is lowest for the ARIMA(0,1,1) model, the difference from the ARIMA(0,1,0) model is not statistically significant (p = .640). The BIC is

```
. corrgram ohhomic

                                            -1       0       1 -1       0       1
 LAG       AC       PAC      Q      Prob>Q  [Autocorrelation]  [Partial Autocor]
---------------------------------------------------------------------------------
1        0.8140   0.8599    33.837   0.0000          |------            |------
2        0.6893   0.1410    58.625   0.0000          |-----             |-
3        0.6133   0.0712    78.684   0.0000          |----              |
4        0.4975  -0.1274    92.185   0.0000          |---              -|
5        0.4394   0.0725   102.96    0.0000          |---               |
6        0.3183  -0.2874   108.75    0.0000          |--              --|
7        0.2085   0.0037   111.3     0.0000          |-                 |
8        0.1094  -0.2122   112.01    0.0000          |                 -|
9        0.0081  -0.0346   112.02    0.0000          |                  |
10      -0.0627  -0.0025   112.27    0.0000          |                  |
11      -0.1815  -0.2930   114.4     0.0000         -|                --|
12      -0.2313   0.0396   117.97    0.0000         -|                  |
13      -0.1739   0.3319   120.04    0.0000         -|                  |--
14      -0.2207  -0.2085   123.48    0.0000         -|                 -|
15      -0.2531  -0.1663   128.14    0.0000        --|                 -|
16      -0.2848  -0.1503   134.22    0.0000        --|                 -|
17      -0.2838   0.0825   140.46    0.0000        --|                  |
18      -0.2461  -0.0532   145.3     0.0000         -|                  |
19      -0.2535  -0.1128   150.62    0.0000        --|                  |
20      -0.2163   0.0502   154.63    0.0000         -|                  |
21      -0.2089  -0.1619   158.51    0.0000         -|                 -|
22      -0.1875   0.1480   161.75    0.0000         -|                  |-
```

Figure 34.4 Correlogram for Ohio homicide data

```
. corrgram wihomic

                                            -1       0       1 -1       0       1
 LAG       AC       PAC      Q      Prob>Q  [Autocorrelation]  [Partial Autocor]
 --------------------------------------------------------------------------------
1        0.8175   0.8698    34.126   0.0000          |------            |------
2        0.7072   0.3045    60.218   0.0000          |-----             |--
3        0.6940   0.3828    85.905   0.0000          |-----             |---
4        0.6615   0.2646   109.77    0.0000          |-----             |--
5        0.6111   0.2146   130.61    0.0000          |----              |-
6        0.4764  -0.2327   143.58    0.0000          |---              -|
7        0.3673  -0.2630   151.48    0.0000          |--              --|
8        0.3506   0.0767   158.85    0.0000          |--                |
9        0.2844  -0.0017   163.83    0.0000          |--                |
10       0.1970   0.0504   166.28    0.0000          |-                 |
11       0.1700   0.3587   168.16    0.0000          |-                 |--
12       0.0825  -0.0750   168.61    0.0000          |                  |
13       0.0017  -0.3926   168.61    0.0000          |               ---|
14       0.0142   0.2822   168.62    0.0000          |                  |--
15      -0.0329   0.0881   168.7     0.0000          |                  |
16      -0.0526   0.0164   168.91    0.0000          |                  |
17      -0.1027  -0.1596   169.73    0.0000          |                 -|
18      -0.1298   0.2479   171.07    0.0000         -|                  |-
19      -0.0936   0.0175   171.8     0.0000          |                  |
20      -0.1077   0.1516   172.79    0.0000          |                  |-
21      -0.1723  -0.3438   175.43    0.0000         -|                --|
22      -0.1594   0.2278   177.78    0.0000         -|                  |-
```

Figure 34.5 Correlogram for Wisconsin homicide data

```
. corrgram ilhomic

                                           -1       0       1 -1       0       1
 LAG       AC       PAC         Q    Prob>Q   [Autocorrelation]  [Partial Autocor]
---------------------------------------------------------------------------------
1        0.8567   0.9150    37.48   0.0000          |------            |-------
2        0.7778   0.2688    69.042  0.0000          |------            |--
3        0.6828  -0.1405    93.909  0.0000          |-----            -|
4        0.6232   0.0595   115.09   0.0000          |----              |
5        0.5526   0.0612   132.14   0.0000          |----              |
6        0.4723  -0.0849   144.88   0.0000          |---               |
7        0.3730   0.0139   153.03   0.0000          |--                |
8        0.2610  -0.1976   157.11   0.0000          |--               -|
9        0.1714  -0.2477   158.92   0.0000          |-                -|
10       0.0671  -0.0837   159.2    0.0000          |                  |
11      -0.0152  -0.0280   159.22   0.0000          |                  |
12      -0.0693   0.0617   159.54   0.0000          |                  |
13      -0.0611   0.4657   159.79   0.0000          |                  |---
14      -0.0835   0.2072   160.29   0.0000          |                  |-
15      -0.1074   0.0351   161.13   0.0000          |                  |
16      -0.1177   0.0269   162.16   0.0000          |                  |
17      -0.1265  -0.0658   163.4    0.0000         -|                  |
18      -0.1058   0.0651   164.3    0.0000          |                  |
19      -0.1128  -0.0542   165.35   0.0000          |                  |
20      -0.0976   0.0111   166.17   0.0000          |                  |
21      -0.1130  -0.3092   167.3    0.0000          |                --|
22      -0.1045   0.2516   168.31   0.0000          |                  |--
```

Figure 34.6 Correlogram for Illinois homicide data

Table 34.2 ARIMA models for homicide rates

Dependent variable: Homicide State	*Statistics/ Coefficients*	*ARIMA (0,1,0)*	*ARIMA (1,0,0)*	*ARIMA (0,1,1)*	*ARIMA (1,1,0)*
Ohio	AIC	127.3998	130.5507	127.1818	127.4642
	BIC	131.1001	136.1643	132.7322	133.0147
	DF	2	3	3	3
	ϕ_1 (p)	–	.877 (.000)*	–	−.199 (.262)
	θ_1 (p)	–	–	−.224 (.156)	–
	constant (p)	.006 (.966)	.587 (.583)	.008 (.949)	.007 (.956)
Wisconsin	AIC	52.5839	54.1320	41.8033	48.7427
	BIC	56.2842	59.7456	47.3537	54.2932
	DF	2	3	3	3
	ϕ_1 (p)	–	.859 (.000)*	–	−.345 (.011)*
	θ_1 (p)	–	–	−.597 (.000)*	–
	constant (p)	.038 (.518)	.068 (.858)	.042 (.051)	.041 (.340)
Illinois	AIC	142.9423	147.2737	141.5559	140.7052
	BIC	146.6426	152.8873	147.1064	146.2556
	DF	2	3	3	3
	ϕ_1 (p)	–	.903 (.000)*	–	−.292 (.034)*
	θ_1 (p)	–	–	−.245 (.073)	–
	constant (p)	.047 (.766)	.808 (.666)	.053 (.658)	.051 (.684)

* $p < .050$

```
. corrgram zohres1

                                            -1       0       1 -1       0       1
 LAG       AC      PAC        Q     Prob>Q   [Autocorrelation]  [Partial Autocor]
---------------------------------------------------------------------------------
1       -0.2027  -0.2027    2.0569  0.1515          -|                 -|
2       -0.0590  -0.1091    2.2353  0.3270           |                  |
3        0.1111   0.0854    2.8819  0.4102           |                  |
4       -0.1234  -0.1026    3.6976  0.4485           |                  |
5        0.2409   0.2438    6.8792  0.2298           |-                 |-
6       -0.1238  -0.0620    7.7405  0.2577           |                  |
7        0.0875   0.1460    8.1812  0.3169           |                  |-
8       -0.0074  -0.0426    8.1844  0.4157           |                  |
9       -0.1297  -0.0860    9.2044  0.4186          -|                  |
10       0.2578   0.1904   13.34    0.2053           |--                |-
11      -0.2066  -0.1412   16.071   0.1385          -|                 -|
12      -0.1836  -0.3199   18.29    0.1072          -|                --|
13       0.1579   0.1497   19.979   0.0957           |-                 |-
14       0.0117   0.1010   19.988   0.1305           |                  |
15       0.0835   0.0591   20.49    0.1539           |                  |
16      -0.1119  -0.1215   21.419   0.1629           |                  |
17      -0.0154  -0.0029   21.438   0.2073           |                  |
18       0.1327   0.1066   22.836   0.1970           |-                 |
19      -0.1527  -0.0430   24.754   0.1688          -|                  |
20       0.1083   0.1698   25.754   0.1741           |                  |-
21      -0.0254  -0.0874   25.812   0.2137           |                  |
```

Figure 34.7 Correlogram for Ohio homicide data

lowest for the ARIMA(0,1,0) model, suggesting that in this instance the AIC may lead us to overparameterize the model. For Wisconsin, however, the random walk with drift represented by the ARIMA(0,1,1) model appears to provide the best fit. It has the lowest AIC and BIC by some considerable margin [and the AIC difference with the ARIMA(0,1,0) model is statistically significant, $p = .001$], and the θ coefficient is statistically significant. Separate testing of ARIMA(1,1,1) and models with higher order autoregressive or moving average components produced statistically nonsignificant coefficients and no improvement in model fit. A third model, this time an ARIMA(1,1,0), appears to provide the best fit for the homicide data from Illinois. The ϕ coefficient is statistically significant in the ARIMA(1,1,0) model [the θ coefficient in the ARIMA(0,1,1) model is not], and both the AIC and the BIC are smallest for this model. Again, testing more complex models produced nonsignificant coefficients and no improvement in fit. For all three models, residuals were computed and their ACFs examined; all autocorrelations were nonsignificant, indicating that the residuals appear to be white noise. This is illustrated in Figure 34.7 for the residuals from the time series analysis of the Ohio data; the residuals for the Wisconsin and Illinois data follow a similar pattern of nonsignificant autocorrelations.

8 Extensions to the simple univariate model

The univariate ARIMA model described above can be extended in several ways, including (a) the incorporation of seasonality and cyclicity into the models; (b) modeling interventions in time series natural experiments or quasi-experimental designs; and (c) addition of time-varying covariates or predictors to the model. *Cyclicity* refers to recurrent patterns that repeat over time, with evenly spaced "peaks" and "valleys" in the series; an example would be the 22-year sunspot cycle (including changes in polarity). *Seasonality* is a more specific form of cyclicity, characteristic of time series such as retail sales and airplane passenger miles, which tend to be higher in some months and lower than others, over a 12-month cycle. Cyclicity in general, and seasonality more specifically, show up as regularly-spaced peaks and valleys in the ACFs and PACFs. They are modeled by adding a seasonal component to the ARIMA model. The standard notation for the seasonal AR, I, and MA components of the model is (P,D,Q), with capital letters for the seasonal components as opposed to lower case letters for the nonseasonal components. The seasonal ARIMA(p,d,q)(P,D,Q) model includes autoregressive, integrated, and moving average parameters for both the seasonal and the nonseasonal components. For example, a seasonal ARIMA$(0,1,0)(0,1,0)_{12}$, where the subscript 12 refers to the *periodicity* of the cyclical or seasonal effect (in this instance, for a seasonal model, 12 months) could be written as $(z_t - z_{t-1}) - (z_{t-12} - z_{t-13}) = (z_t - z_{12}) - (z_{t-1} - z_{t-13}) = z_t - z_{t-1} - z_{t-12} + z_{t-13} = a_t$. This involves differencing for both the first order integrated $I(d) = I(1)$ effects $(z_t - z_{t-1})$ and the parallel $(z_{t-12} - z_{t-13})$, and also the seasonal $I(D) = I(1)$ effects $(z_t - z_{t-12})$ and the parallel $(z_{t-1} - z_{t-13})$. Autoregressive and moving average effects are similarly incorporated into the model. As a practical matter, cyclic or seasonal models can be incorporated into the ARIMA model and estimated using the same software as for a nonseasonal ARIMA model. As such, they represent an extension to the ARIMA model that is methodologically relatively easy to incorporate, but is significant in its substantive importance. An alternative to ARIMA modeling of cyclic patterns is spectral analysis, as described in detail by Jenkins and Watts (1968) and more briefly by, e.g., Wei (2005); see also Wei, Chapter 36, in this volume.

8.1 Intervention analysis

In quasi-experimental research (e.g., Shadish et al., 2002), the simple time series design is represented as a series of measurements of the outcome, split into two segments by the insertion of an intervention or treatment; e.g., as illustrated below, for a time series of 12 total observations, half occurring before and half after the treatment or intervention, we can express the series as

$$O_1O_2O_3O_4O_5O_6 \text{ X } O_7O_8O_9O_{10}O_{11}O_{12}$$

where O represents an observation, the subscript on O indexes the time at which the observation was taken, and X indicates the treatment or intervention, here occurring between the sixth and seventh observations. Both the intervention and, separately, the effects of the intervention, may be either discrete, occurring at a specific time and not before or after that time, or it may be persistent, absent prior to some specific time and present after that time. An example of a discrete intervention, occurring between two observations and at no other time, might be a one-time increase in funding to provide equipment to a law enforcement agency or a short-term therapy regimen. An example of a persistent intervention might be a change in policy, for example presumptive arrest in domestic violence cases, or an ongoing, perhaps permanent treatment regimen, for example medication to suppress the effects of human immunodeficiency virus (HIV). Regardless of whether the treatment itself is discrete or persistent, the *effects* of the treatment may be either discrete or persistent, and if persistent, persistent over a shorter or a longer span of time, possibly decaying more or less rapidly over time. Further details on intervention models may be found in Box and Jenkins (1970; see also the more recent edition by Box et al., 1994), McCleary and Hay (1980), Wei (2006), and Yaffee and McGee (2000). In the present volume, further discussion of time series intervention models is deferred to Chapter 36 by Sanders and Ward.

8.2 Multivariate time series analysis

Intervention analysis focuses on one-time discrete or persistent effects of interventions that effectively split a time series into “before” and “after” segments. Other influences on time series cannot be represented so simply, and may themselves be time series varying over time in much the same way as the outcome. Bivariate and multivariate time series models are available to explain an outcome time series in terms of one or more input time series, and to assess whether it is more plausible that the purported outcome time series is, in fact, dependent upon the purported input time series, or whether the evidence offers better support for the purported outcome’s being predictive of the purported input. There are several approaches to the analysis of time series involving several variables, either with or without an *a priori* definition of which variables are outcomes or predictors, several of which (autoregressive maximum likelihood models, lagged endogenous variable models, Box-Jenkins ARIMA models) are described and compared in Chapter 36 by Sanders and Ward in this volume. Here, therefore, only a brief overview is provided. There are other models, particularly as used in econometrics, which can be applied to the analysis of time series; for further details on distributed lag models, Kalman filter and state space models, and related topics, see, e.g., Amemiya (1985) or Johnston and DiNardo (1997); for a briefer treatment of time series regression techniques, see Ostrom (1990).

8.3 Autoregressive error models

As described in, e.g., Ostrom (1990; see also Sanders and Ward, Chapter 36, in this volume), the use of ordinary least squares (OLS) multiple regression analysis for multivariate time series analysis has the serious disadvantage that the errors tend not to be independent, but are instead themselves autocorrelated. This results in underestimation of standard errors, overestimation of explained variance, and increased risk of Type I (falsely rejecting the null hypothesis) error. One approach to this problem is to explicitly model the autoregressive nature of the errors. This may be done by assuming an AR(1) or higher order AR(p) model and estimating the model using maximum likelihood or other estimation techniques. Statistical software for estimating the AR(p) model in the autoregressive error model context and the ARIMA model context should produce the same results if the same estimation method (e.g., the same maximum likelihood estimation algorithm) is used. The autoregressive error model may be extended to also model heteroscedasticity in the error variance, using *autoregressive conditional heteroscedasticity (ARCH)* or *generalized autoregressive conditional heteroscedasticity (GARCH)* models (for introductory treatments of which see, e.g., Wei, 2006; Yaffee and McGee, 2000). In the ARCH model, assuming a contemporaneous effect of X on Z (for a lagged effect, the subscript t would be replaced by t−1 for the predictors $x_{k,t}$) $z_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + e_t$, where e_t is assumed to be normally distributed with mean zero and variance $h_t = \alpha_0 + \alpha_1(e_{t-1})^2 + \alpha_2(e_{t-2})^2 + \cdots + \alpha_p(e_{t-p})^2$ in which case the residual variance (and by implication the error variance) is not constant but (a) varies over time and (b) depends on prior values of the time-specific residuals. This variation is explicitly modeled in the ARCH model. In the GARCH model, h_t depends not only on past values of e_t but also on past values of h_t. The difference of the simple autoregressive error model from the ARCH and GARCH models lies in the explicit modeling of not only the value but also the variance of e_t in the ARCH and GARCH models.

8.4 Granger causality

An approach that tests for both causal direction and strength of causal influence was proposed by Granger (1969). For two variables Z_t and Y_t, both of which can be expressed as stationary time series with zero means,

$$z_t = \sum_{j=1}^{m1} a_j z_{t-j} + \sum_{j=1}^{m2} a_j y_{t-j} + e_t$$

and

$$y_t = \sum_{j=1}^{m3} c_j z_{t-j} + \sum_{j=1}^{m4} d_j y_{t-j} + f_t$$

where e_t and f_t are taken to be two uncorrelated white noise series, and m is greater than zero but less than the length of the time series. According to the criterion of Granger causality, Y causes Z if some b_j is statistically significantly different from zero. Correspondingly, Z causes Y if some c_j is statistically significantly different from zero. In effect, the question posed by the test for Granger causality becomes, “Is there variation in one variable which cannot be explained by past values of that variable, but can be explained by past values of another variable?” If the answer is yes, then the second variable “Granger-causes” the first. Instantaneous effects (e.g., from x_2 to y_2) are excluded from the model. One issue here is the choice of m1, m2, m3, and m4, the number of lags to include in each part of the model. In general, the more prior values of the endogenous variable are in the equation, the greater is the likelihood of rejecting the hypothesis of Granger causality, but the inclusion of additional values of the endogenous variable may have no significant effect beyond some number. This number may be estimated by modeling the endogenous variable as an autoregressive time series, or by calculating separate ordinary least-squares regression models and examining the change in the explained variance (R^2) produced by the inclusion of each additional lagged endogenous variable (e.g., by the addition of y_{t-4}). If there is no statistically significant change in the explained variance, there would seem to be little point in including this term in the equation. If only a single lag is used for each outcome (m1 = 1 and m4 = 1), we have for both X and Y the lagged endogenous variable (LEV) OLS model described by Sanders and Ward in Chapter 36. Granger causality may be examined in its own right, or as part of the analysis of dynamic regression/linear transfer function and vector autoregression models, as described below.

8.5 Transfer function, dynamic regression/linear transfer function, and vector autoregression models

Transfer function models are typically explicit in identifying which variable is an outcome or effect and which is a predictor or cause. In a transfer function model, whether X causes Y or Y causes X may be based on *a priori* knowledge of the process, but can also be assessed based on the *cross-correlation function* between X and Y, which examines correlations of y_t with $x_{t-1}, x_{t-2}, \ldots, x_{t-m3}$ and the correlations of x_t with $y_{t-1}, y_{t-2}, \ldots, y_{t-m2}$ (here using m2 and m3 to indicate the lags examined, the same as in the Granger causality equations above). In examining the cross-correlation functions, it is assumed that the respective series are, or have been rendered, stationary. Once the model has been identified, based on the cross-correlation function, the impact of X on Y (or Y on X) is estimated using maximum likelihood techniques, as in estimation of simple univariate ARIMA models. In order to calculate the impact of the predictor on the outcome, it is necessary to *prewhiten* the series by (a) using information about the ARIMA process generating the input series to transform the input series to a white noise process, then (b) applying the same set of transformations to the outcome series, before (c) calculating the impact of the input series on the outcome series. For example, if the input series can be represented as

$x_t = a_t + x_{t-1} + \phi_1(x_{t-1} - x_{t-2})$ [an ARIMA(1,1,0) process], we calculate $a_t = x_t - x_{t-1} - \phi_1(x_{t-1} - x_{t-2})$ and apply the same transformation to Y to obtain the output series, now designated Z: $z_t = y_t - y_{t-1} - \phi_1(y_{t-1} - y_{t-2})$. Prewhitening removes any spurious correlation based on trend between X and Z (the transformed series Y). Prewhitening is relatively easily achieved for bivariate transfer function models but, for models with two or more inputs, it can be more complicated.

In an alternative approach called *dynamic regression (DR)* or *linear transfer function (LTF)* modeling (see, e.g., Yaffee and McGee, 2000), the argument is made that there is no need for prewhitening, as long as the input series are not highly correlated. The approach here is similar to causal modeling using OLS multiple regression, with the outcome and predictor variables identified based on theory rather than empirically, using a cross-correlation function, although a Granger causality test provides useful insight into the identification of which variables are predictors and which are outcomes. In the DR/LTF approach, the input and outcome series are rendered stationary, the error term is modeled as an autoregressive process (or possibly an integrated process if ϕ_1 is approximately equal to 1), and the outcome series Z is modeled as a function of one or more lagged values of predictor series $X_1, X_2, \ldots, X_K$. The process for modeling DR/LTF models does require a series of steps to make sure that what is modeled is not spurious correlation or random variation, but these steps are easily performed in the context of the basic ARIMA modeling approach.

Vector autoregressive (VAR) models (e.g., Brandt and Williams, 2007; Stock and Watson, 2001; Wei, 2006) take a more exploratory approach to modeling multivariate time series. The emphasis here, as described by Brandt and Wilson, is on avoiding restrictive and quite possibly false assumptions made in alternative methods (ARIMA transfer function modeling, structural equation modeling, autoregressive error models), and instead letting the data determine the structure of the model as much as possible. In a VAR model, all of the variables are treated as endogenous, subject to influence by all of the other variables in the model. VAR models allow for the existence of simultaneous and mutual influence among the variables in the model, but the modeling strategy itself allows analysis using OLS multiple regression calculating separate equations for each of the variables in the model. In a two variable model, the equations

$$z_t = \eta_1 y_t + \Sigma\rho_{1,p} z_{t-p} + \Sigma\gamma_{1,p} y_{t-p} + u_{1,t}$$

and

$$y_t = \eta_2 z_t + \Sigma\rho_{2,p} y_{t-p} + \Sigma\gamma_{2,p} z_{t-p} + u_{2,t}$$

can be reparameterized as

$$z_t = \alpha_1 + \Sigma\phi_{1,p} z_{t-p} + \Sigma\beta_{1,p} y_{t-p} + \varepsilon_{1,t}$$

and

$$y_t = \alpha_2 + \Sigma\phi_{2,p} y_{t-p} + \Sigma\beta_{2,p} z_{t-p} + \varepsilon_{2,t}$$

where ϕ and β can be expressed in terms of the original parameters (and the α coefficients are just constants added to the equation, consistent with the usual practice in OLS regression). The reparameterized equations can be estimated using OLS regression, because they are not expressed in terms of simultaneous effects of one endogenous variable on another (hence simultaneity bias in the original but not the reparameterized equations).

As described in Brandt and Williams (2007), the purposes of VAR models are assessing causal relationships, for which purpose Granger causality analysis is performed as part of the analysis; assessing dynamic impacts, which is done by inverting the VAR equations to find the moving average representation of the model; and determining how much of the variation of each variable is attributable to dynamic changes in other variables, with particular interest in how much of an impact the contemporaneous influences have compared to the lagged influences. In VAR models, in contrast to both transfer function and DR/LTF models, differencing is discouraged, because it obscures information of interest about trends in the model. Also in contrast to transfer function models, prewhitening is not appropriate in VAR models because it can alter the dynamics of the model and change the assessment of Granger causality. The principal advantages of VAR lie in its ease of computation, its avoidance (compared to other multivariate time series methods) of potentially false assumptions used to identify the structure of the model, and its ability to investigate a potentially broad range of possible specifications of the dynamic relationships in the data. Criticisms have been that it cannot truly assess causal relationships, that it is atheoretical, and that VAR models tend to be overparameterized in their attempt to avoid simplifying assumptions that might result in more parsimonious models.

9 Forecasting

Thinking in the context of longitudinal research, with an implicit interest in description and explanation, possibly including theory testing, time series analysis is one of many tools, having utility for addressing some specific problems that may arise in longitudinal research. This is not, however, entirely consistent with the context in which time series analysis techniques were developed. Time series analysis, perhaps more than other techniques of analysis used in longitudinal research, is preeminently an applied technique. The underlying reason for the interest in describing time series processes is to predict future values of the outcome variable, in order to control the outcome by controlling inputs. *Forecasting* future values of the time series is a relatively minor topic in the context of longitudinal research, where greater emphasis is placed on explaining past outcomes than on accurately predicting, and intervening to control, future outcomes. In the context of time series analysis, however, forecasting and control (the subtitle of Box and Jenkins, 1976) may be the whole point of time series analysis. A similar applied orientation is characteristic of intervention analysis using time series techniques; the whole point is to determine how much of an impact over what span of time we can expect as a result of an intervention. In keeping with the longitudinal research focus of this volume, forecasting is treated only briefly here, but greater detail is available in most (particularly book-length) treatments of time series analysis.

One approach to forecasting derives from the curve fitting approach described at the beginning of this chapter. One simply uses the formula for the fitted curve, with future time points as input, to forecast the future value of the series. The danger in this, of course, is that the curve which has been fitted to the observed data may to some extent be capitalizing on random variation in the data, and predicted values may quickly diverge from future values of the outcome that have not been used to model the data. Consider once again the pattern in Figure 34.2. At the very end of the series, the predicted value is increasing, and because the curve at the end is quickly dominated by the cubic and quartic terms in the polynomial equation, the curve will continue to increase rapidly; but the right tail of the observed data actually appears to have a downward trend. Other methods of forecasting such as exponential smoothing (see, e.g., Yaffee and McGee, 2000) include the use of long (e.g, 10 time points) moving averages to smooth the curve, coupled with one-step forecasts that take the existing moving average at time t, forecast the next value of the outcome at time $t+1$, calculate a new moving average of the same length that includes the time $t+1$ forecast as the last data point in the moving average, then use that moving average (including the forecast data point) to forecast the next

$(t+2)$ value of the series. This one-step-ahead approach can include adjustments for trends in the data to improve the forecasts, and is an alternative to forecasting based solely on a deterministic model for trend.

As indicated in the ARIMA analysis of the data in Figure 34.2, the trend is stochastic drift, the result of random differences between adjacent values of the outcome plus the lingering effect of a past random shock, and in this sense, the curve fitted to the data really is capitalizing on random variation. Similarly, the moving average plus exponential smoothing approach has limited information upon which to base its forecast. Using the ARIMA model to forecast the series also holds little promise for improvement. Because the process is driven by random shocks whose expected value is $E(a_t) = 0$, the "best" forecast in this instance may simply be to forecast that the value of the next point in the series will be equal to the previous value of the series. In short, when the process involves only random shocks, there is little real information upon which to base a forecast other than the most recent value of the series. When the values of the series depend on more than one past value of the series, more accurate forecasting may be possible. To the extent that there are deterministic trends, seasonality, and cyclicity in the data, this information can also be used to produce more accurate forecasts. Different methods can also be combined to improve forecasts.

One way to assess the accuracy of a forecasting model is to split the data into two sets, a longer series of earlier observations on which the forecast model is developed, and a shorter series of later observations on which the model can be tested. Based on a review (Yaffee and McGee, 2000) of comparative studies of forecasting accuracy, including simulation results, ARIMA models appear to do well for short-term forecasting, while simple regression techniques may outperform other methods for medium- and long-term forecasts. This is because in the short term, random variation (which is modeled in the Box-Jenkins ARIMA approach) may have a substantial impact, as may seasonality, but in the longer term, random variation and seasonality tend to be dominated by the other systematic components of the process, particularly deterministic trends; it is the processes that are primarily driven by random shocks that are most difficult to forecast accurately. Also as noted by Yaffee and McGee (2000), combined methods, particularly when they include the Box-Jenkins ARIMA approach, generally outperform single-method forecasting approaches. Unsurprisingly, the farther in the future the forecast is made, the less accurate it will be with any methods, and the more information (as opposed to random variation) is available on the process, the more accurate the forecasts will be.

10 Conclusion

Time series analysis may be used to answer any of the following questions.

(1) What is the functional relationship of an outcome variable Z with time? Here the answer is an equation with Z as the dependent variable, or some function of Z as the dependent variable, where the function of Z may involve subtraction of the mean to center the series at zero, differencing to remove linear or higher order trends, taking the natural logarithm (sometimes done to stabilize the variance over the length of the series), or other transformations. The predictors are functions of time, for example expressing Z as a polynomial function of time, $z_t = \lambda_0 + \lambda_1 t + \lambda_2 t^2 + \cdots + \lambda_M t^M$; or an exponential function, $z_t = \lambda_0 + \lambda_1 e^t$; or some other function of time. This question may best be answered by graphical representation and curve fitting techniques. The objective may or may not be to describe the process, but it is to describe the pattern.

(2) How does the current value of an outcome variable Z depend on past values of Z? Here the answer is an equation with Z (or

some function of Z) as the dependent variable and one or more lagged values of Z as predictors, possibly with an assumed coefficient of 1, as in the ARIMA(0.1,0) random walk or the ARIMA(0,1,1) random walk with drift models; or with estimated ϕ coefficients where $z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p}$; or some combination of the two. This question may best be answered by an ARIMA(p,d,q) model with $d + p > 0$.

(3) How do past random shocks affect the current value of an outcome Z? Here Z is expressed as a function of the current random shock a_t and one or more past random shocks $a_{t-1}, a_{t-2}, \ldots, a_{t-q}$ such that $z_t = \theta_0 + a_t, + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}$, and for a series centered on its mean, θ_0 will be equal to zero. This question may best be answered by an ARIMA(p,d,q) model with $q > 0$.

(4) How do past (and possibly contemporaneous) values of one or more time-varying predictors $Y_{k,t}$ affect the value of the outcome variable Z? Here, $Y_{k,t}$ may be a one-time intervention, coded zero for all but the one time in which the intervention occurs and one for the single point in time of the intervention, for an acute, nonenduring intervention; or it may be coded zero for the time prior to the intervention and one for the time subsequent to the intervention, for an intervention with enduring effects; or it may have some other coding, corresponding to the hypothesized duration of the effects; or the duration itself may be an unknown to be modeled. Alternatively, or in the same analysis, $X_{j,t}$ may be another time series that influences the outcome. Alternatives for answering this question include autocorrelated error models, autoregression with correction for heteroscedasticity (ARCH) and generalized autoregression with correction for heteroscedasticity (GARCH) models, multivariate lagged endogenous variable models, transfer function models, dynamic regression (DR)/linear transfer function (LTF) models, and vector autoregression (VAR) models.

(5) We may want to ask all of these questions at the same time; the result would be a complex model, $\nabla^d z_t = a_t + \theta_0 + \sum \theta_q a_{t-q} + \sum \phi_p z_{t-p} + \sum \gamma_m f_m(t) + \sum \beta_k y_{k,t.}$ In principle, it should be possible to estimate such a model. In practice, trends would not be represented by both differencing and some function of time, and either the analysis would proceed on the residuals of a deterministic trend model represented by $\sum \gamma_m f_m(t)$, in which case the explicit time component $\sum \gamma_m f_m(t)$ would be subtracted from z_t and the integrated component ∇^d would be excluded from the model; or the model would drop the explicit time component $\sum \gamma_m f_m(t)$ and retain the integrated component ∇^d; leaving either the model with the residualized deterministic time trend $[z_t - \sum \gamma_m f_m(t)] = a_t + \theta_0 + \sum \theta_q a_{t-q} + \sum \phi_p z_{t-p} + \sum \beta_k y_{k,t}$, or the model $\nabla^d z_t = a_t + \theta_0 + \sum \theta_q a_{t-q} + \sum \phi_p z_{t-p} + \sum \beta_k y_{k,t}$. Once again, for a time series centered on its mean, θ_0 would drop out of the model. Because it readily models all of these different components, the ARIMA model, possibly incorporating the DR/LTF approach for time-varying covariates, provides perhaps the most flexible approach to time series modeling, as long as there are sufficient data (typically 50 or more time-specific observations) to support the use of the model. For shorter series, other approaches may be preferable, and for large numbers of cases, alternative models such as latent or multilevel growth curve models are better suited to model the relationships of predictors to outcome variables.

Software

Time series analysis can be performed with specialized software such as RATS, or with modules in existing general-purpose statistical packages such as SAS, SPSS, Stata, and SYSTAT. For time series curve fitting, the SPSS CURVEFIT routine is particularly useful, while Stata has user-friendly options for ARCH and GARCH models. For the time series analysis performed here, SPSS and Stata statistical routines were used.

Bibliographic note

[1]Several time series texts and associated statistical software manuals have been used in the preparation of this chapter. Box and Jenkins (1970) is the classic on ARIMA time series analysis, and Box et al. (1994) is a more recent edition of this classic. The parallel classic for spectral analysis is Jenkins and Watts (1968). An out-of-print and slightly out-of-date but well written introductory treatment of ARIMA time series analysis can be found in McCleary and Hay (1980). A more up-to-date introduction from a social science perspective is Yaffee and McGee (2000), which includes detailed instruction on the use of SAS and SPSS for time series analysis. A brief introduction with an emphasis on the comparison among SAS, SPSS, and SYSTAT time series analysis routines for ARIMA time series models is offered by Tabachnick and Fidell (2007) on their website, www.ablongman.com/tabachnick5e. Wei (2006) offers a more detailed and more advanced treatment of time series analysis, including spectral analysis.

References

Amemiya, T. (1985). *Advanced Econometrics.* Cambridge, MA: Harvard University Press.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control,* 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

Brandt, P. T. and Williams, J. T. (2007). *Multiple Time Series Models.* Thousand Oaks, CA: Sage.

Cromwell, J. B., Hannan, M. J., Labys, W. C. and Terraza, M. (1994). *Multivariate Tests for Time Series Models.* Thousand Oaks: Sage.

Cromwell, J. B., Labys, W. C. and Terraza, M. (1994). *Univariate Tests for Time Series Models.* Thousand Oaks: Sage.

Hamblin, R. L., Jacobsen, R. B. and Miller, J. L. L. (1973). *A Mathematical Theory of Social Change.* New York: Wiley.

Jenkins, G. M. and Watts, D. G. (1968) *Spectral Analysis and Its Applications.* San Francisco: Holden-Day.

Johnston, J. and DiNardo, J. (1997). *Econometric Methods,* 4th edn. New York: McGraw-Hill.

Kohfeld, C. W. and Decker, S. H. (1990). Time series, panel design, and criminal justice: A multistate, multiwave design. In K. L. Kempf (ed.), *Measurement Issues in Criminology.* New York: Springer-Verlag.

McCleary, R. and Hay, R. A., Jr. (1980). *Applied Time Series Analysis for the Social Sciences.* Beverly Hills, CA: Sage.

Ostrom, C. W., Jr. (1990). *Time Series Analysis: Regression Techniques.* Thousand Oaks, CA: Sage.

Scheffé, H. (1959). *The Analysis of Variance.* New York: Wiley.

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Stock, J. H. and Watson, M. W. (2001). Vector autoregression. *Journal of Economic Perspectives,* 5: 101–115.

Tabachnick, B. G. and Fidell, L. S. (2007). *Using Multivariate Statistics,* 5th edn. Boston: Allyn and Bacon.

Wei, W. E. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods,* 2nd edn. Boston: Addison-Wesley.

Yaffee, R., with McGee, M. (2000). *Introduction to Time Series Analysis and Forecasting with Applications of SAS and SPSS.* San Diego: Academic Press.

Chapter 35

Spectral analysis

William W.S. Wei

Spectral analysis is a statistical method used to analyze a time series dataset identifying statistically important frequencies present in a time series to see whether it contains periodic or cyclical components. After reviewing some basic time series concepts, we introduce periodogram analysis that is a useful technique to search for hidden periodicities. We then introduce the spectrum of a time series, a sample spectrum and its smoothing. In the process, we introduce some commonly used spectral and lag windows and a procedure to obtain confidence intervals for the underlying spectral ordinates. We also discuss cross-spectrum that can be used to analyze the relationships between frequency components in two time series. Empirical examples are used to illustrate the concepts and procedures. The chapter ends with some mathematical justifications that have been used throughout the chapter.

1 Introduction

A time series is often referred to as an ordered sequence of observations. The ordering is usually through time, especially in terms of some equally spaced time intervals. For example, we observe daily calls to directory assistance, monthly international airline passengers, quarterly unemployment numbers, annual imports and exports, and various mortality rates and crime rates. The body of statistical methodology available for studying time series is referred to as *time series analysis*. More formally, a time series is a realization of a *time series process* that is a family of time indexed random variables, Z_t, where t belongs to an index set. For most of our discussion, we assume that the index set is the set of all integers. The fundamental goal of time series analysis is to investigate the underlying process through an observed time series or realization. Thus, it is important to understand some fundamental characteristics of a time series process.

A time series process is said to be *strictly stationary* if the joint distribution of $(Z_{t_1}, \ldots, Z_{t_n})$ is the same as the joint distribution of $(Z_{t_1+k}, \ldots, Z_{t_n+k})$ for any n-tuple $(t_1, \ldots, t_n)$ and k of integers. The terms *strongly stationary* and *completely stationary* are also used to denote a strictly stationary process. Since strict stationarity is in terms of its distribution function, and it is very difficult or impossible to verify a general distribution function, for a given time series process Z_t, $t = 0, \pm 1, \pm 2, \ldots,$ we often concentrate our study on some of its important parameters such as moments including the *mean function* of the process

$$\mu_t = E(Z_t) \tag{1.1}$$

the *variance function* of the process

$$\sigma_t^2 = Var(Z_t) = E(Z_t - \mu_t)^2 \tag{1.2}$$

the *autocovariance function* between Z_{t_1} and Z_{t_2}

$$\gamma(t_1, t_2) = E(Z_{t_1} - \mu_{t_1})(Z_{t_2} - \mu_{t_2}) \tag{1.3}$$

and the *autocorrelation function* between Z_{t_1} and Z_{t_2}

$$\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sqrt{\sigma^2_{t_1}}\sqrt{\sigma^2_{t_2}}} \tag{1.4}$$

For a strictly stationary process, since the distribution function is the same for all t, the mean function, $\mu_t = \mu$, is a constant, provided $E(|Z_t|) < \infty$. Similarly, if $E(Z_t^2) < \infty$, then the variance function, $\sigma_t^2 = \sigma^2$, is constant, and the autocovariance and autocorrelation functions depend only on the time difference, i.e., $\gamma(t, t + k) = E(Z_t - \mu)(Z_{t+k} - \mu) = \gamma_k$ and $\rho(t, t+k) = \gamma_k/\gamma_0 = \rho_k$, where we note that $\gamma_0 = Var(Z_t) = \sigma^2$.

A time series process is said to be *nth order weakly stationary* if all its joint moments up to order n exist and are time invariant, i.e., independent of time origin. Therefore, a second-order weakly stationary process will have a constant mean and variance, with the autocovariance and the autocorrelation functions being functions of the time difference alone. The terms *weakly stationary* or *stationary in the wide sense* or *covariance stationary* are also used to describe a second-order weakly stationary process.

It is noted that process and model are sometimes interchangeably used. Hence a time series model is also used to refer to a time series process. The simplest time series process or model is a *white noise process* e_t that is a sequence of uncorrelated random variables from a fixed distribution with constant mean $E(e_t) = \mu_e$, usually assumed to be 0, constant variance $Var(e_t) = \sigma_e^2$ and $\gamma_k = Cov(e_t, e_{t+k}) = 0$ for all $k \neq 0$.

A time series process is said to be a normal or *Gaussian process* if its joint distribution is normal. Because a normal distribution is uniquely characterized by its first two moments, strictly stationary and weakly stationary are equivalent for a Gaussian process. Unless otherwise mentioned, the time series processes that we discuss are assumed to be Gaussian.

2 A simple periodic model and harmonic analysis

Consider a simple zero mean time series Z_t that exhibits a periodic or cyclical pattern with a *fundamental period* N, which is the smallest time period for this repetitive pattern to hold. For this periodic process, both the time series Z_t and its autocovariance function γ_k will exhibit spikes at multiple lags of N, i.e., at times $t = jN$ and lags $k = jN$ for $j = \pm 1, \pm 2, \ldots$. A natural representation of this periodic or cyclical phenomenon is the following simple sinusoidal model:

$$Z_t = \mu + \alpha\cos(\omega t + \phi) + e_t \tag{2.1}$$

where e_t is a zero mean Gaussian white noise process, α is the amplitude or height of the cycle, ϕ is the phase or location of cycle peak relative to time origin zero, and $\omega = 2\pi/N$ is the *fundamental frequency* corresponding to the given fundamental period N. More often the following equivalent form that is more convenient for computing parameter estimates is used:

$$Z_t = \mu + a\cos(\omega t) + b\sin(\omega t) + e_t \tag{2.2}$$

where

$$\alpha = \sqrt{a^2 + b^2} \tag{2.3}$$

and

$$\phi = \tan^{-1}(-b/a) \tag{2.4}$$

Given the observations Z_t for $t = 1, \ldots, N$, the least square estimates are given by

$$\hat{\mu} = \frac{1}{N}\sum_{t=1}^{N} Z_t \tag{2.5}$$

$$\hat{a} = \frac{2}{N}\sum_{t=1}^{N} Z_t\cos(\omega t) \tag{2.6}$$

and

$$\hat{b} = \frac{2}{N}\sum_{t=1}^{N} Z_t \sin(\omega t) \tag{2.7}$$

The amplitude and phase of the cycle can then be estimated from equations (2.3) and (2.4), respectively, using (2.6) and (2.7).

The least square estimates from (2.5) to (2.7) follow from the trigonometric results shown in Section 7 that:

$$\sum_{t=1}^{N} \cos(\omega t) = \sum_{t=1}^{N} \sin(\omega t) = 0 \tag{2.8}$$

$$\sum_{t=1}^{N} \cos(\omega t)\sin(\omega t) = 0 \tag{2.9}$$

$$\sum_{t=1}^{N} \cos(\omega t)\cos(\omega t) = N/2 \tag{2.10}$$

and

$$\sum_{t=1}^{N} \sin(\omega t)\sin(\omega t) = N/2 \tag{2.11}$$

Using the well known result from regression analysis that

$$\sum_{t=1}^{N} (Z_t - \bar{Z})^2 = \sum_{t=1}^{N} (Z_t - \hat{Z}_t)^2 + \sum_{t=1}^{N} (\hat{Z}_t - \bar{Z})^2 \tag{2.12}$$

i.e., total sum of squares = sum of squares due to error + sum of squares due to regression, we obtain the sum of squares due to the cyclical or periodic component as

$$\sum_{t=1}^{N} \left[\hat{a}\cos(\omega t) + \hat{b}\sin(\omega t)\right]^2 = N(\hat{a}^2 + \hat{b}^2)/2 \tag{2.13}$$

which follows from equations (2.9), (2.10), and (2.11). This is proportional to the squared amplitude of the fitted sinusoid, and it is the amount of variance accounted for by the periodic component at frequency ω. Corresponding to the model in (2.2), this sum of squares due to regression has two degrees of freedom, one each for a and b, and by comparing with the sum of squares due to error that has $(N-3)$ degrees of freedom, we can use the standard F test to assess the significance of this periodic component. The detecting procedure introduced above is often known as *harmonic analysis.*

3 Periodogram analysis

In most situations, the exact periods of cycles are not known, and we want to identify them using available time series Z_t for $t = 1, \ldots, N$. A very natural approach is to extend the harmonic analysis of Section 2 at all frequencies $2\pi k/N$ for $k = 0, 1, \ldots, N/2$. That is, for a given time series of N observations, we consider the following representation:

$$Z_t = \sum_{k=0}^{[N/2]} (a_k \cos(\omega_k t) + b_k \sin(\omega_k t) \tag{3.1}$$

where $\omega_k = 2\pi k/N, k = 0, 1, \ldots, [N/2]$, and $[x]$ is the greatest integer less than or equal to x. Suppose that the trigonometric sine and cosine functions $\cos(\omega_k t)$ and $\sin(\omega_k t)$ are defined on a finite number of N points, i.e., for $t = 1, 2, \ldots, N$, the system

$$\{\cos(\omega_k t), \sin(\omega_k t) : k = 0, 1, \ldots, [N/2]\} \tag{3.2}$$

contains exactly N nonzero functions, which follow from the sine function being identically zero for $k = 0$ and $k = [N/2]$ if N is even. Moreover, as shown in Section 7, the system is a set of *orthogonal functions*, i.e.,

$$\sum_{t=1}^{N} \cos(\omega_k t)\cos(\omega_j t) = \begin{cases} N, \quad k = j = 0 \quad \text{or } N/2 \ (N \text{ even}) \\ N/2, \ k = j \neq 0 \quad \text{or } N/2 \ (N \text{ even}) \\ 0, k \neq j \end{cases} \tag{3.3}$$

$$\sum_{t=1}^{N} \sin(\omega_k t)\sin(\omega_j t) = \begin{cases} 0, \quad k = j = 0 \text{ or } N/2 \ (N \text{ even}) \\ N/2, \ k = j \neq 0 \text{ or } N/2 \ (N \text{ even}) \\ 0, k \neq j \end{cases} \tag{3.4}$$

and

$$\sum_{t=1}^{N} \cos(\omega_k t)\sin(\omega_j t) = 0 \text{ for all } k \text{ and } j. \tag{3.5}$$

Thus, the system in (3.2) forms a basis, and equation (3.1) fits the N data points exactly. The representation is known as the *Fourier series* of Z_t, and $\omega_k = 2\pi k/N, k = 0, 1, \ldots, [N/2]$, are called *Fourier frequencies*. Using the results in (3.3), (3.4), and (3.5), we immediately obtain the following:

$$a_k = \begin{cases} \dfrac{1}{N}\sum\limits_{t=1}^{N} Z_t \cos(\omega_k t), \ k = 0 \text{ and} \\ \quad k = N/2 \text{ if } N \text{ is even} \\ \dfrac{2}{N}\sum\limits_{t=1}^{N} Z_t \cos(\omega_k t) \\ \quad k = 1, 2, \ldots, [(N-1)/2] \end{cases} \tag{3.6}$$

$$b_k = \frac{2}{N}\sum_{t=1}^{N} Z_t \sin(\omega_k t)$$

$$k = 1, 2, \ldots, [(N-1)/2] \tag{3.7}$$

which are known as *Fourier series coefficients*. They are, in fact, essentially the least squares estimates of the coefficients in fitting the following regression model:

$$Z_t = \sum_{k=0}^{[N/2]} a_k \cos(\omega_k t) + b_k \sin(\omega_k t) + e_t \tag{3.8}$$

Note that equation (2.2) is a special case where we consider the components only at frequencies ω_0 and ω_1, and the component at ω_0 gives $a_0 = \sum_{t=1}^{N} Z_t/N$ that is actually the mean of the series. The fitting will be perfect. Multiplying Z_t on both sides of (3.1), summing from $t = 1$ to $t = N$, and using the relation (3.6) and (3.7), we have

$$\sum_{t=1}^{N} Z_t^2 = \begin{cases} Na_0^2 + \dfrac{N}{2}\sum\limits_{k=1}^{[(N-1)/2]} (a_k^2 + b_k^2) \\ \quad \text{if } N \text{ is odd,} \\ Na_0^2 + \dfrac{N}{2}\sum\limits_{k=1}^{[(N-1)/2]} (a_k^2 + b_k^2) + Na_{N/2}^2 \\ \quad \text{if } N \text{ is even.} \end{cases} \tag{3.9}$$

Hence,

$$\sum_{t=1}^{N} (Z_t - \overline{Z})^2 = \begin{cases} \dfrac{N}{2}\sum\limits_{k=1}^{[(N-1)/2]} (a_k^2 + b_k^2) \text{ if } N \text{ is odd} \\ \dfrac{N}{2}\sum\limits_{k=1}^{[(N-1)/2]} (a_k^2 + b_k^2) \\ \quad + Na_{N/2}^2 \text{ if } N \text{ is even} \end{cases} \tag{3.10}$$

and the result is presented as the following analysis of variance in Table 35.1.

Table 35.1 Analysis of variance table for periodogram analysis

Source	*Degrees of freedom*	*Sum of squares*
Frequency $\omega_0 = 0$ (Mean)	1	Na_0^2
Frequency $\omega_1 = 2\pi/N$	2	$\frac{N}{2}(a_1^2 + b_1^2)$
Frequency $\omega_2 = 4\pi/N$	2	$\frac{N}{2}(a_2^2 + b_2^2)$
$\vdots$	$\vdots$	$\vdots$
Frequency $\omega_{[(N-1)/2]}$ $= [(N-1)]\,2\pi/N$	2	$\frac{N}{2}\left\{a_{[(N-1)/2]}^2 + b_{[(N-1)/2]}^2\right\}$
Frequency $\omega_{N/2} = \pi$ (exists only for even N)	1	$Na_{N/2}^2$
Total	N	$\sum_{t=1}^{N} Z_t^2$

The quantity $I(\omega_k)$ defined by

$$I(\omega_k) = \begin{cases} Na_0^2 \ k=0 \\ \frac{N}{2}(a_k^2+b_k^2) \\ \quad k=1,\ldots\ [(N-1)/2] \\ Na_{N/2}^2 \ k=N/2 \text{ when } N \text{ is even} \end{cases} \tag{3.11}$$

is called the *periodogram*, and the procedure is known as *periodogram analysis*. It was introduced by Arthur F. Schuster (1898), who used the technique to disprove C. G. Knott's claim of periodicity in earthquake occurrences. Schuster (1906) went on to apply the method to analyzing annual sunspot activity and found the approximate eleven-year cycle of the sunspot series.

Assume that Z_t for $t=1,\ldots,$ and N are i.i.d. $N(0,\sigma^2)$. We have

$$E(a_k) = \frac{2}{N}\sum_{t=1}^{N} E(Z_t)\cos(\omega_k t) = 0$$

and

$$Var(a_k) = \frac{4}{N^2}\sum_{t=1}^{N}\sigma^2[\cos(\omega_k t)]^2$$
$$= \frac{4\sigma^2}{N^2}\sum_{t=1}^{N}[\cos(\omega_k t)]^2 = \frac{4\sigma^2}{N^2}\frac{N}{2} = \frac{2\sigma^2}{N}$$

where we use the fact that a_k and a_j are independent for $k \neq j$ because of (3.3). Hence, the a_k for $k=1,2,\ldots,[(N-1)/2]$, are i.i.d. $N(0,2\sigma^2/N)$, and the $Na_k^2/2\sigma^2$ for $k=1,2,\ldots,[(N-1)/2]$, are i.i.d. chi-squares with one degree of freedom. Similarly, the $Nb_k^2/2\sigma^2$ for $k=1,2,\ldots,[(N-1)/2]$, are i.i.d. chi-squares with one degree of freedom. Furthermore, $Na_k^2/2\sigma^2$ and $Nb_j^2/2\sigma^2$ for all k and $j=1,2,\ldots,[(N-1)/2]$, are independent, because a_k and b_j are normal and by the orthogonal property of (3.2), we have

$$Cov(a_k,b_j) = \frac{4}{N^2}E\left[\sum_{t=1}^{N} Z_t\cos(\omega_k t), \sum_{u=1}^{N} Z_u\sin(\omega_j u)\right]$$
$$= \frac{4}{N^2}\sum_{t=1}^{N}[E(Z_t^2)\ \cos(\omega_k t)\cdot\sin(\omega_j t)]$$
$$= \frac{4\sigma^2}{N^2}\sum_{t=1}^{N}[\cos(\omega_k t)\cdot\sin(\omega_j t)]$$
$$= 0 \text{ for any } k \text{ and } j \tag{3.12}$$

It follows that

$$\frac{I(\omega_k)}{\sigma^2} = \frac{N}{2\sigma^2}(a_k^2+b_k^2) \tag{3.13}$$

for $k=1,2,\ldots,[N/2]$ are i.i.d. chi-squares with two degrees of freedom, denoted as $\chi^2(2)$. Clearly, $[I(0)/\sigma^2] = Na_0^2$ when $k=0$ and $[I(\pi)/\sigma^2] = Na_{N/2}^2$ when $k=[N/2]$ and N is even are each a chi-square distribution with one degree of freedom. With the adjustment for $I(\pi)$ in mind, we assume, without loss of generality, that the length of the series N is odd in the remaining discussion.

Clearly, if in Table 35.1 the only significant component sum of squares is at frequency ω_1, then model (3.1) reduces to model (2.2), and the sum of squares for all the remaining frequencies will be combined to become the sum of squares for error with $(N-3)$ degrees of freedom. Hence, as indicated earlier in Section 2, to test the significance of the periodic component at ω_1, we can use the following test statistic

$$F = \frac{[N(a_1^2+b_1^2)/2\sigma^2]/2}{\left[\sum_{j=2}^{[N/2]} N(a_j^2+b_j^2)/2\sigma^2\right]/(N-3)}$$
$$= \frac{(N-3)N(a_1^2+b_1^2)}{2\left[\sum_{j=2}^{[N/2]} N(a_j^2+b_j^2)\right]}$$

In other words, to test for the significance of the component at frequency ω_k, we use the test statistic

$$F = \frac{[N(a_k^2 + b_k^2)/2]/2}{\left[\sum_{j=1, j\neq k}^{[N/2]} N(a_j^2 + b_j^2)/2\right]/(N-3)}$$

$$= \frac{(N-3)(a_k^2 + b_k^2)}{2\left[\sum_{j=1, j\neq k}^{[N/2]} (a_j^2 + b_j^2)\right]} \qquad (3.14)$$

which follows the F-distribution with 2 and $(N-3)$ degrees of freedom, denoted as $F(2, N-3)$. Since the period P and frequency ω are related by

$$P = 2\pi/\omega \qquad (3.15)$$

once a significant frequency ω_k is found, the period or the length of the cycle is found to be $2\pi/\omega_k$. More generally, we can test whether a series contains multiple m periodic components by postulating the model

$$Z_t = \mu + \sum_{i=1}^{m} (a_{k_i} \cos(\omega_{k_i} t)$$
$$+ b_k \sin(\omega_k t) + e_t \qquad (3.16)$$

where the e_t are i.i.d. $N(0, \sigma^2)$, $\omega_{k_i} = 2\pi k_i/N$, and the set $I = \{k_i : i = 1, \ldots, m\}$ is a subset of $\{k : k = 1, \ldots, [N/2]\}$. The corresponding test statistic will be

$$\mathrm{F} = \frac{(N-2m-1)\left[\sum_{i=1}^{m} (a_{k_i}^2 + b_{k_i}^2)\right]}{2m\left[\sum_{j=1, j\notin I}^{[N/2]} (a_j^2 + b_j^2)\right]} \qquad (3.17)$$

which follows the F-distribution with $2m$ and $(N-2m-1)$ degrees of freedom, i.e., $F(2m, N-2m-1)$.

4 Tests for hidden periodic components

In practice, even if we believe that a time series contains a periodic component, the underlying frequency is often unknown. For example, we might test the null hypothesis $H_0 : \alpha = \beta = 0$ against the alternative $H_1 : \alpha \neq 0$ or $\beta \neq 0$ in the model

$$Z_t = \mu + \alpha \cos(\omega t) + \beta \sin(\omega t) + e_t, \qquad (4.1)$$

where e_t is a Gaussian white noise $N(0, \sigma^2)$ process, and the frequency ω is unknown. Because the frequency ω is unknown, the F-distribution and the test statistics as discussed in Section 3 are not directly applicable. The periodogram analysis, however, is still useful. In fact, the original purpose of the periodogram was to search for hidden periodicities. If the time series indeed contains a single periodic component at frequency ω, it is hoped that the periodogram $I(\omega_k)$ at the Fourier frequency ω_k closest to ω will be the maximum. Thus, we can search out the maximum period ordinate and test whether this ordinate can be reasonably considered as the maximum in a random sample of $[N/2]$ i.i.d. random variables, each being a multiple of a chi-square distribution with two degrees of freedom. Thus, a natural test statistic will be

$$I^{(1)}(\omega_{(1)}) = \max\{I(\omega_k) : k = 1, \ldots, [N/2]\} \qquad (4.2)$$

where $\omega_{(1)}$ is used to denote the Fourier frequency with the maximum periodogram ordinate.

In 1929, Ronald A. Fisher derived an exact test for $I^{(1)}(\omega_{(1)})$ based on the following statistic

$$T = \frac{I^{(1)}(\omega_{(1)})}{\sum_{k=1}^{[N/2]} I(\omega_k)} \qquad (4.3)$$

Under the null hypothesis of a $N(0, \sigma^2)$ white noise process for Z_t, Fisher (1929) showed that

$$P(T > g) = \sum_{j=1}^{r} (-1)^{(j-1)} \binom{K}{j} (1 - jg)^{K-1} \qquad (4.4)$$

where $K = (N-1)/2$ if N is odd and $K = (N/2-1)$ if N is even, $g > 0$, and r is the largest integer less than $1/g$. Thus, for any given significance level α, we can use equation (4.4) to find the critical value g_α such that $P(T > g_\alpha) = \alpha$. We will reject the null hypothesis and conclude that the series contains a periodic component if the T value calculated from the series is larger than g_α. This test is known as Fisher's test. The critical values of T for the significance level of $\alpha = .05$ as given by Fisher is shown in Table 35.2. As shown in the third column of Table 35.2, for most practical purposes, equation (4.4) can be approximated with the first term, i.e.,

$$P(T > g) \approx K(1-g)^{K-1} \tag{4.5}$$

and this is useful for the case when K is not listed on the table.

For the model in (4.1), a significant value of T leads to the rejection of the null hypothesis and implies that there exists a periodic component in the series at some frequency ω. This frequency, however, is not necessarily equal to $\omega_{(1)}$, because $\omega_{(1)}$ is chosen only from the Fourier frequencies and not from all possible frequencies between 0 and π.

Table 35.2 The critical values of Fisher's test for the maximum periodogram ordinate at $\alpha = .05$

K	g_α	g_α *(by the first term only)*
5	.68377	.68377
10	.44495	.44495
15	.33462	.33463
20	.27040	.27046
25	.22805	.22813
30	.19784	.19794
35	.17513	.17525
40	.15738	.15752
45	.14310	.14324
50	.13135	.13149

$K = (N-1)/2$ if N is odd and $K = (N/2-1)$ if N is even.

Herman O. Hartley (1949), however, has shown that the unknown ω with the maximum periodic component can be safely estimated by $\omega_{(1)}$ since $P\left(\left|\omega - \omega_{(1)}\right| > 2\pi/N\right) < \alpha$, the significance level of the test.

Let $I^{(2)}(\omega_{(2)})$ be the second largest periodogram ordinate at Fourier frequency $\omega_{(2)}$. Peter Whittle (1952) suggested that we could extend Fisher's test for the second largest ordinate based on the test statistic

$$T_2 = \frac{I^{(2)}(\omega_{(2)})}{\sum_{k=1}^{[N/2]} I(\omega_k) - I^{(1)}(\omega_{(1)})} \tag{4.6}$$

where the distribution in (4.4) is taken as the distribution of T_2 with K being replaced by $(K-1)$. The procedure can be continued until an insignificant result is reached. It leads to an estimate of m, the number of periodic components present in the series.

Before going further into the topic, let us consider an example of using the method to analyze a dataset. It should be noted, however, that from the above discussion the length of a time series determines the Fourier frequencies and the periods to be tested. It is important that the length of a time series used in periodogram analysis should be an integer multiple of the fundamental period or cycle length that we are trying to detect. Otherwise, an artifact called *leakage* may occur, where the variance of a real cyclical component that cannot be accurately detected spills over into the sum of squares of other frequencies that are detected by the periodogram analysis. For example, if the fundamental period of a series is 12, there is no leakage if the length of 24 observations is used. If 23 observations are used, the periodogram analysis detects the components at frequencies $2\pi k/23, k = 1, \ldots, 11$, which no longer include a component of period 12.

There are many statistical packages available to perform the spectral analysis introduced in this chapter. They include SAS issued by SAS Institute (2003), S-Plus issued by Insightful

Corporation (2005), SPSS issued by SPSS, Inc. (2006), and many others. They produce very similar results. We will use SAS to perform all our analyses in this chapter.

Example 1

We illustrate the periodogram analysis using the classical series of the monthly totals Y_t of international airline passengers (in thousands) from January 1949 to December 1960 quoted by Robert G. Brown (1963) and made popular by George E. P. Box and Gwilym M. Jenkins (1976). The data are plotted in Figure 35.1. The series exhibits a clear periodic behavior with higher peaks in late summer months and secondary peaks in the spring.

Because of its clear nonconstant variance, we calculate the periodogram of the natural logarithms of the series as plotted in Figure 35.2. The components at low frequencies near zero are clearly dominating. This is due to the clear upward trend shown in the series.

The dominating components at low frequencies due to nonstationarity often make other components insignificant. To see the fine details of underlying characteristics, it is important to reduce a nonstationary series to a stationary series. There are many methods available to remove nonstationarity in a series. One may consider differencing or trend removal. In this example, we will analyze the series of monthly growth rates defined by

$$Z_t = 100\,[\log(Y_t) - \log(Y_{t-1})] \qquad (4.7)$$

Since the data is a monthly series, we are interested in detecting a possible seasonal period of 12 months. As noted earlier, to avoid possible leakage, we will consider observations from December 1949 to December 1960 to get the monthly growth rates of a full 11 years. The periodogram of these growth rates is plotted in Figure 35.3. The fine details of the periodic phenomenon are now evident. For the significance test at $\alpha = .05$ with the number of observations $N = 132$, we have $F_{.05}(2, 129) \approx 3$. The only significant components occur at the fundamental frequency $2\pi/12$ and its harmonics $2\pi k/12, k = 1, 2, 3, 4, 5$, and 6, indicating a strong periodic phenomenon with a fundamental period of 12 months. Thus, the monthly growth rates of international airline passengers can be very well presented by the cyclical model:

$$Z_t = \mu + \sum_{k=1}^{6} (a_k \cos(\omega_k t) + b_k \sin(\omega_k t) + e_t \qquad (4.8)$$

where the e_t are i.i.d. $N(0, \sigma^2)$, and $\omega_k = 2\pi k/12$.

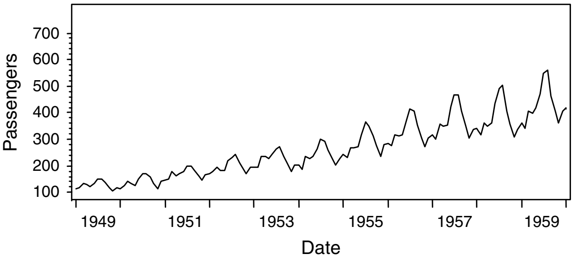


Figure 35.1 Monthly totals (in thousands) of international airline passengers between January 1949 and December 1960

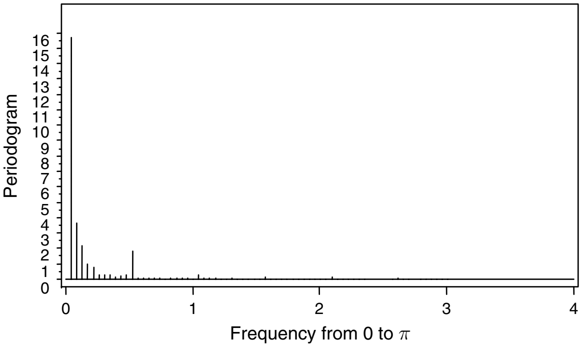


Figure 35.2 Periodogram of the natural logarithms of monthly totals of international airline passengers

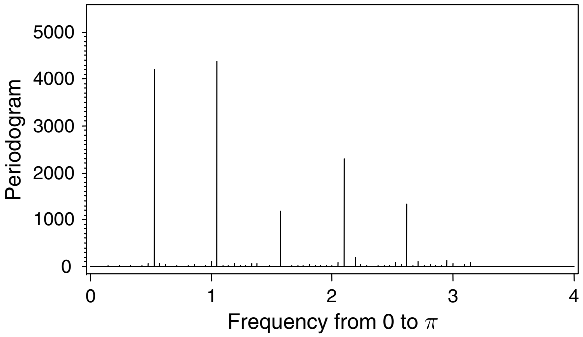


Figure 35.3 Periodogram of the monthly growth rates of international airline passengers

5 The spectrum of time series and its estimation

Consider a stationary time series Z_t with the autocovariance function γ_k. The *autocovariance generating function* $\gamma(B)$ is defined as

$$\gamma(B) = \sum_{k=-\infty}^{\infty} \gamma_k B^k \tag{5.1}$$

where the variance of the process γ_0 is the coefficient of $B^0 = 1$ and the autocovariance of lag k, γ_k, is the coefficient of both B^k and B^{-k}. If the autocovariance sequence γ_k is absolutely summable, i.e., $\sum_{k=-\infty}^{\infty} |\gamma_k| < \infty$, then the *spectrum* or the *spectral density* exists and equals

$$f(\omega) = \frac{1}{2\pi}\gamma(e^{-i\omega}) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_k e^{-i\omega k} \tag{5.2}$$

where $-\pi \leq \omega \leq \pi$. Note that because $\gamma_k = \gamma_{-k}$, $\sin(0) = 0$, $\sin(-\omega k) = -\sin(\omega k)$, and $\cos(-\omega k) = \cos(\omega k)$, the spectrum (5.2) can be written equivalently as

$$f(\omega) = \frac{1}{2\pi}\left[\gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \cos(\omega k)\right] \tag{5.3}$$

Note that the spectrum $f(\omega)$ is a continuous real-valued nonnegative function. Furthermore, because $f(\omega) = f(-\omega)$, it is a symmetric even function, and its graph is normally presented only for $0 \leq \omega \leq \pi$. For a given time series of N observations, however, the autocovariance of the maximum lag that can be calculated is $\hat{\gamma}_{(N-1)}$. Thus, we estimate $f(\omega)$ by

$$\begin{aligned}\hat{f}(\omega) &= \frac{1}{2\pi}\sum_{-(N-1)}^{(N-1)} \hat{\gamma}_k e^{-i\omega k} \\ &= \frac{1}{2\pi}\left[\hat{\gamma}_0 + 2\sum_{k=1}^{N-1} \hat{\gamma}_k \cos(\omega k)\right]\end{aligned} \tag{5.4}$$

and call it the *sample spectrum*.

To examine the properties of the sample spectrum, let us consider $\hat{f}(\omega_j)$ at the Fourier frequency $\omega_j = 2\pi j/N, j = 1, \ldots, [N/2]$. At these Fourier frequencies, the sample spectrum and the periodogram are closely related. To see this, we note that

$$\begin{aligned}I(\omega_j) &= \frac{N}{2}(a_j^2 + b_j^2) \\ &= \frac{N}{2}(a_j - ib_j)(a_j + ib_j) \\ &= \frac{N}{2}\left[\frac{2}{N}\sum_{t=1}^{N} Z_t(\cos(\omega_j t) - i\sin(\omega_j t)\right] \\ &\quad \times \left[\frac{2}{N}\sum_{t=1}^{N} Z_t(\cos(\omega_j t) + i\sin(\omega_j t)\right] \\ &= \frac{2}{N}\left[\sum_{t=1}^{N} Z_t e^{-i\omega_j t}\right]\left[\sum_{t=1}^{N} Z_t e^{i\omega_j t}\right] \\ &= \frac{2}{N}\left[\sum_{t=1}^{N} (Z_t - \bar{Z})e^{-i\omega_j t}\right]\left[\sum_{t=1}^{N} (Z_t - \bar{Z})e^{i\omega_j t}\right] \\ &= \frac{2}{N}\sum_{t=1}^{N}\sum_{s=1}^{N} (Z_s - \bar{Z})(Z_t - \bar{Z})e^{-i\omega_k(t-s)}\end{aligned} \tag{5.5}$$

where we use the Euler relation, $e^{\pm i\omega_j} = \cos(\omega_j) \pm i\sin(\omega_j)$, and the fact that for $j \neq 0$,

$$\sum_{t=1}^{N} e^{\pm i\omega_j t} = \sum_{t=1}^{N} \cos(\omega_j t) \pm i \sin(\omega_j t)$$
$$= \sum_{t=1}^{N} \cos(\omega_0 t) \cos(\omega_j t)$$
$$\pm\, i \sum_{t=1}^{N} \cos(\omega_0 t) \sin(\omega_j t) = 0$$

Now, $\hat{\gamma}_k = \frac{1}{N} \sum_{t=k+1}^{N} (Z_{t-k} - \bar{Z})(Z_t - \bar{Z})$
Let $k = t - s$ in (5.5), we have

$$I(\omega_j) = 2 \sum_{k=-(N-1)}^{(N-1)} \frac{1}{N} \sum_{t=k+1}^{N} (Z_{t-k} - \bar{Z})(Z_t - \bar{Z}) e^{-i\omega_j k}$$
$$= 2 \sum_{k=-(N-1)}^{(N-1)} \hat{\gamma}_k e^{-i\omega_j k}$$
$$= 2 \left(\hat{\gamma}_0 + 2 \sum_{k=1}^{(N-1)} \hat{\gamma}_k \cos(\omega_j k) \right) \tag{5.6}$$

Hence, from (5.4), we have

$$\hat{f}(\omega_j) = \frac{1}{4\pi} I(\omega_j) \tag{5.7}$$

Since the cosine function is a periodic continuous function with a period 2π, the periodogram in (5.6) can easily be extended to all ω as the following periodic continuous function between $-\pi$ to π:

$$I(\omega) = 2 \sum_{k=-(N-1)}^{(N-1)} \hat{\gamma}_k e^{-i\omega k}$$
$$= 2 \left(\hat{\gamma}_0 + 2 \sum_{k=1}^{(N-1)} \hat{\gamma}_k \cos(\omega k) \right)$$

Because $I(\omega_k)$ for $\omega_k = 2\pi k/N, k = 1, \ldots, [N/2]$ is the standard output from a periodogram analysis, equation (5.7) becomes a natural candidate for estimating the spectrum and is known as the *periodogram estimator of the spectrum.*

To see the properties of this estimator, we recall from (3.13) that for a Gaussian white noise series with mean 0 and constant variance σ^2, $\hat{f}(\omega_k)$, for $k = 1, \ldots, [N/2]$, are actually distributed independently and identically as $(1/4\pi)\sigma^2\chi^2(2) = (\sigma^2/2\pi)\chi^2(2)/2$, which is to be denoted as

$$\hat{f}(\omega_k) \sim \frac{\sigma^2}{2\pi} \frac{\chi^2(2)}{2} \tag{5.8}$$

where we note that from (5.3), $\sigma^2/2\pi$ is in fact the spectrum for the given Gaussian white noise process with mean 0 and variance σ^2. More generally, following the same arguments, for a general Gaussian process with a spectrum $f(\omega)$, the sample spectrum calculated at Fourier frequencies ω_k in (5.7) has the following asymptotic distribution

$$\lim_{N\to\infty} \hat{f}(\omega_k) = f(\omega_k) \frac{\chi^2(2)}{2} \tag{5.9}$$

Hence,

$$\lim_{N\to\infty} E(\hat{f}(\omega_k)) = f(\omega_k) \tag{5.10}$$

and

$$\lim_{N\to\infty} Var(\hat{f}(\omega_k)) = Var\left[f(\omega_k) \frac{\chi^2(2)}{2} \right]$$
$$= [f(\omega_k)]^2 \tag{5.11}$$

which is independent of the sample size N. Thus, although the periodogram estimator of the spectrum in (5.7) calculated at a Fourier frequency is asymptotically unbiased, it is an unsatisfactory estimator because it is not consistent. The variance of $\hat{f}(\omega_k)$ does not reduce to zero as the sample size N goes to infinity. The effort of correcting this deficiency leads to the smoothing of the periodogram estimator.

5.1 The smoothed periodogram estimator

A natural way to reduce the variance of the periodogram estimator of the spectrum is to smooth the periodogram locally in the neighborhood of the target frequency. In other words, we obtain a smoothed periodogram estimator from the following weighted average of m values to the right and left of a target frequency ω_k, i.e.,

$$\hat{f}_W(\omega_k) = \sum_{j=-m}^{m} W_N(\omega_j)\hat{f}(\omega_k - \omega_j) \tag{5.12}$$

where m is a function of N, which is often chosen such that $m \to \infty$ but $(m/N) \to 0$ as $N \to \infty$; and $W_N(\omega_j)$ is the *weighting function* with the following properties

$$\sum_{j=-m}^{m} W_N(\omega_j) = 1 \tag{5.13}$$

$$W_N(\omega_j) = W_N(-\omega_j) \tag{5.14}$$

and

$$\lim_{N\to\infty} \sum_{j=-m}^{m} W_N^2(\omega_j) = 0 \tag{5.15}$$

The weighting function $W_N(\omega_j)$ is called the *spectral window* because only some of the spectral ordinates are utilized and shown in the smoothing. If the $f(\omega)$ is flat and constant within the window, then

$$\lim_{N\to\infty} E[\hat{f}_W(\omega_k)] = \lim_{N\to\infty} \sum_{j=-m}^{m} W_N(\omega_j)E[\hat{f}(\omega_k - \omega_j)$$
$$= f(\omega_k)$$

and

$$\lim_{N\to\infty} Var\left[\hat{f}_W(\omega_k)\right]$$
$$= [f(\omega_k)]^2 \lim_{N\to\infty} \sum_{j=-m}^{m} W_N^2(\omega_j) = 0 \tag{5.16}$$

where we use the result that the periodogram ordinates at different Fourier frequencies are independent.

The property (5.16) of the spectral window implies that the variance of the smoothed periodogram estimator decreases as N and hence m increases. The values of m represent the number of frequencies used in the smoothing. This value is directly related to the width of the spectral window, also known as the *bandwidth* of the window. As the bandwidth increases, more spectral ordinates are averaged; hence, the resulting estimator becomes smoother, more stable, and has smaller variance. Unless $f(\omega)$ is really flat, however, the bias also increases as the bandwidth increases because more and more spectral ordinates are used in the smoothing. We are thus forced to compromise between variance reduction and bias, a common dilemma with many statistical estimators.

Because the periodogram is periodic with period 2π, when the window covers frequencies that fail to lie entirely in the range between $-\pi$ to π, we can extend the periodogram using this periodic property. Equivalently, we can fold the weights back into the interval $-\pi$ to π. As the periodogram is also symmetric about frequency zero, calculation is only necessary for the frequency range between zero and π. Also, as shown earlier, because the periodogram at frequency zero reflects the sample mean of the series and not the spectrum, it is not included in the smoothing, and the value at ω_1 is used in its place.

From (5.4), we see that an alternative approach to perform the smoothing is to apply a weighting function $\lambda_N(k)$ to the sample autocovariances, i.e.,

$$\hat{f}_W(\omega) = \frac{1}{2\pi} \sum_{k=-(N-1)}^{(N-1)} \lambda_N(k)\hat{\gamma}_k e^{-i\omega k} \tag{5.17}$$

Because the sample autocovariance function $\hat{\gamma}_k$ is symmetric and $\hat{\gamma}_k$ is less reliable for larger k, the weighting function $\lambda_N(k)$ should be chosen to be symmetric with its weights inversely proportional to the magnitude of k. Thus,

$$\hat{f}_W(\omega) = \frac{1}{2\pi} \sum_{k=-M}^{M} \lambda_N(k)\hat{\gamma}_k e^{-i\omega k} \tag{5.18}$$

where the weighting function $\lambda_N(k)$ is chosen to be an absolutely summable sequence

$$\lambda_N(k) = \lambda\left(\frac{k}{M}\right) \tag{5.19}$$

which is often derived from a bounded continuous function $\lambda(x)$ satisfying

$$|\lambda(x)| \le 1$$
$$\lambda(0) = 1$$
$$\lambda(x) = \lambda(-x)$$
$$\lambda(x) = 0, |x| > 1$$

The value of M is the *truncation point* that depends on the sample size N. This weighting function $\lambda_N(k)$ for the autocovariances is called the *lag window*. It can be shown (see William W.S. Wei, 2006, p. 305) that the spectral window and lag window form a Fourier transform pair. The spectral window is the Fourier transform of the lag window, i.e.,

$$W_N(\omega) = \frac{1}{2\pi} \sum_{k=-M}^{M} \lambda_N(k) e^{-i\omega k} \tag{5.20}$$

and the lag window is the inverse Fourier transform of the spectral window, i.e.,

$$\lambda_N(k) = \int_{-\pi}^{\pi} W_N(\omega) e^{i\omega k} d\omega,\ k = 0, \pm 1, \ldots, \pm M \tag{5.21}$$

Both the terms spectral window and lag window were introduced by Ralph B. Blackman and John W. Tukey (1958). The weighing function was the standard term used in the earlier literature.

There are many windows introduced in the literature. The following are some commonly used windows.

Rectangular window:

$$\lambda_N(k) = \begin{cases} 1, |k| \le M \\ 0, |k| > M \end{cases} \tag{5.22a}$$

where M is the truncation point less than $(N-1)$, derived from the continuous rectangular function

$$\lambda(x) = \begin{cases} 1, |x| \le 1 \\ 0, |x| > 1 \end{cases} \tag{5.22b}$$

The corresponding spectral window is given by

$$W_N(\omega) = \frac{1}{2\pi} \frac{\sin[\omega(M+1/2)]}{\sin(\omega/2)} \tag{5.22c}$$

Bartlett window: Maurice S. Bartlett (1950) proposed the lag window

$$\lambda_N(k) = \begin{cases} 1 - |k|/M, |k| \le M \\ 0, |k| > M \end{cases} \tag{5.23a}$$

based on the triangular function

$$\lambda(x) = \begin{cases} 1 - |x|, & |x| \le 1 \\ 0, & |x| > 1 \end{cases} \tag{5.23b}$$

and hence the window is also known as the triangular window. The corresponding spectral window is given by

$$W_N(\omega) = \frac{1}{2\pi M} \left[\frac{\sin(\omega M/2)}{\sin(\omega/2)}\right]^2 \tag{5.23c}$$

Blackman-Tukey window: Blackman and Tukey (1958) proposed the lag window

$$\lambda_N(k) = \begin{cases} 1 - 2a + 2a\cos(\pi k/M), & |k| \le M \\ 0, & |k| > M \end{cases} \tag{5.24a}$$

based on the continuous function

$$\lambda(x) = \begin{cases} 1-2a+2a\cos(\pi x), & |x| \le 1 \\ 0, & |x| > 1 \end{cases} \tag{5.24b}$$

where $0 < a \le .25$. The corresponding spectral window is given by

$$\begin{aligned} W_N(\omega) = {} & \frac{a}{2\pi}\,\frac{\sin[(\omega-\pi/M)(M+1/2)]}{\sin[(\omega-\pi/M)/2]} \\ & + \frac{(1-2a)}{2\pi}\,\frac{\sin[\omega(M+1/2)]}{\sin(\omega/2)} \\ & + \frac{a}{2\pi}\,\frac{\sin\{(\omega+\pi/M)(M+1/2)]}{\sin[(\omega+\pi/M)/2]} \end{aligned} \tag{5.24c}$$

When $a = .23$, the window is known as Hamming or Tukey-Hamming window, and when $a = .25$, the window is also known as Hanning or Tukey-Hanning window or Tukey window.

Parzen window: Emanuel Parzen (1961) suggested the lag window

$$\lambda_N(k) = \begin{cases} 1-6(k/M)^2+6(|k|/M)^3, & |\mathrm{k}| \le M/2 \\ 2(1-|k|/M)^3, & M/2 < |k| \le M \\ 0, & |k| > M \end{cases} \tag{5.25a}$$

based on the continuous function

$$\lambda(x) = \begin{cases} 1-6x^2+6|x|^3, & |x| \le 1/2 \\ 2(1-|x|)^3, & 1/2 < |x| \le 1 \\ 0, & |x| > 1 \end{cases} \tag{5.25b}$$

The corresponding spectral window is given by

$$\begin{aligned} W_N(\omega) = {} & \frac{3}{8\pi M^3}\left\{\frac{\sin(\omega M/4)}{[\sin(\omega/2)]/2}\right\}^4 \\ & \times\left\{1-2[\sin(\omega/2)]^2/3\right\} \end{aligned} \tag{5.25c}$$

In addition, because of its simplicity, especially in terms of deriving the sampling properties of spectrum estimates, a simple $M-$ term moving average of the periodogram surrounding a target Fourier frequency ω is also commonly used. Most statistical packages also accept a user-provided weighting function. The spectral windows given in (5.22c) through (5.25c) are obtained through (5.20). For details, we refer readers to Wei (2006, Section 13.3.3).

The quality of a smoothed sample spectrum is determined by the shape of the window and the bandwidth of the window. The spectrum estimates for the same window shape and different bandwidth are different. In smoothing a sample spectrum, we are concerned not only about the design of a spectral window with a desirable shape known as *window carpentry* as Tukey called it, but also about the bandwidth of a window. The latter concern is often more crucial and difficult in spectral analysis because for a given window shape there is no single criterion for choosing the optimal bandwidth. To ease the difficulty, the following steps are often suggested. First, choose a spectral window with a desirable shape. Initially calculate spectral estimates using a large bandwidth and then recalculate the estimates using gradually smaller bandwidths until the required stability and resolution are achieved. This procedure is often known as "*window closing*." Alternatively, because the bandwidth of a spectral window is inversely related to the truncation point M used in the lag window, the bandwidth can also be determined by choosing a *truncation point* M such that $\hat{\gamma}_k$ for $k > M$ are negligible.

Example 2

For illustration, we now obtain the spectrum estimate by smoothing the periodogram of the monthly growth rates of international airline passengers using the Parzen window with $M = 5$. The resulting spectrum is plotted in Figure 35.4. The smooth curve retains a clear periodic phenomenon with a fundamental period of 12 shown in a monthly seasonal time series.

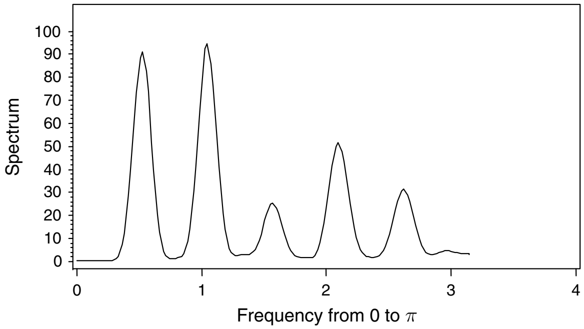


Figure 35.4 A smoothed sample of the monthly growth rates of international airline passengers

5.2 Approximate confidence interval for the spectrum

Consider a time series from a process with the spectrum $f(\omega)$. Let $\hat{f}(\omega_k)$ be the unsmoothed sample spectral ordinate at Fourier frequencies $\omega_k = 2\pi k/N$, with $\omega_k \neq 0$ or π, we have from (5.9) that they are independently and identically distributed as

$$\hat{f}(\omega_k) \sim f(\omega_k)\frac{\chi^2(2)}{2} \tag{5.26}$$

Or equivalently,

$$\frac{2\hat{f}(\omega_k)}{f(\omega_k)} \sim \chi^2(2) \tag{5.27}$$

However, this result is no longer true for a general smoothed sample spectrum. Let $\hat{f}_W(\omega)$ be the general smoothed sample spectral ordinate at frequency ω. John W. Tukey (1949) suggested approximating the distribution of $\hat{f}_W(\omega)$ by a distribution of the form $c\chi^2(\nu)$, where c and ν are chosen so that its mean and variance are equal to the asymptotic mean and variance of $\hat{f}_W(\omega)$. Thus, it can be shown (see William W.S. Wei, 2006, p. 316) that asymptotically we have

$$\frac{\nu\hat{f}_W(\omega)}{f(\omega)} \sim \chi^2(\nu) \tag{5.28}$$

where ν is known as *equivalent degree of freedom for the smoothed spectrum*. The value of ν depends on the window used in the smoothing and is computed as

$$\nu = \frac{2N}{M\int_{-1}^{1}\lambda^2(x)dx} \tag{5.29}$$

where $\lambda(x)$ is the continuous weighting function used in the associated lag window. For example, for Bartlett window, we have

$$\begin{aligned}\nu &= \frac{2N}{M\int_{-1}^{1}(1-|x|)^2dx}\\ &= \frac{2N}{M[\int_{-1}^{0}(1+x)^2dx+\int_{0}^{1}(1-x)^2dx]} = \frac{3N}{M}\end{aligned}$$

Thus, from (5.28), we obtain the following $(1-\alpha)100\,\%$ confidence interval for the spectrum

$$\frac{\nu\hat{f}_W(\omega)}{\chi^2_{\alpha/2}(\nu)} \leq f(\omega) \leq \frac{\nu\hat{f}_W(\omega)}{\chi^2_{1-\alpha/2}(\nu)} \tag{5.30}$$

where $\chi^2_{\alpha}(\nu)$ is the upper α% point of the chi-square distribution with ν degrees of freedom.

Example 3

For illustration, we calculate the 95% confidence interval for the spectrum of the underlying process that generates the monthly growth rates of international airline passengers using the Parzen window with $M=5$. For $N=132$ and $M=5$, we have, from equations (5.25b) and (5.29), $\nu = 3.709N/M = 3.709(132/5) = 97.9176 \approx 98$. Now $\chi^2_{.975}(98) = 72.501$ and $\chi^2_{.025}(98) = 127.282$, the 95% confidence interval for $f(\omega)$, from (5.30), becomes

$$.77\hat{f}_W(\omega) \leq f(\omega) \leq 1.35\hat{f}_W(\omega)$$

where $\hat{f}_W(\omega)$ is the spectrum estimate using the Parzen window with $M=5$ given in Figure 35.4. For example, since $\hat{f}_W(.5236) = 90.9371$, the

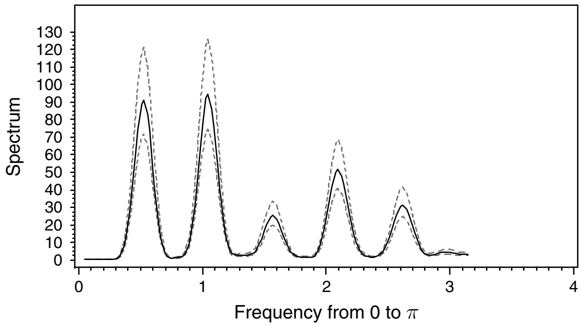


Figure 35.5 The 95% confidence intervals for the spectrum of the monthly growth rates of the international airline passengers using the Parzen window

95% confidence interval for $f(\omega)$ at $\omega = .5236$ is given by

$$70.022 \leq f(\omega = .5236) \leq 122.765$$

The confidence intervals (dotted lines) for other frequencies can be calculated similarly, and they are shown in Figure 35.5.

6 Relationships between two times series and cross-spectrum

6.1 Cross-covariance and cross-spectrum

Very often time series are observed concurrently on two variables, and we are interested in detecting and describing the relationships between two series. Given two processes X_t and Y_t for $t = 0, \pm 1, \pm 2, \ldots$, they are said to be jointly stationary if X_t and Y_t are each stationary and the cross-covariance between X_t and Y_t is a function of time difference only. In such a case, the *cross-covariance function* between X_t and Y_t is given by

$$\gamma_{XY}(k) = E\left[(X_t - \mu_X)(Y_{t+k} - \mu_Y)\right] \qquad (6.1)$$

for $k = 0, \pm 1, \pm 2, \ldots$, where $\mu_X = E(X_t)$ and $\mu_Y = E(Y_t)$. Upon standardization, we obtain the *cross-correlation function*:

$$\rho_{XY}(k) = \frac{\gamma_{XY}(k)}{\sigma_X \sigma_Y} \qquad (6.2)$$

where σ_X and σ_Y are the standard deviation of X_t and Y_t, respectively. While the cross-covariance and cross-correlation functions are useful measures of the relations between two processes, they are only in terms of integer lags. If we want to study the relationship between the two series at any frequency and hence at any time lag, we will extend the univariate spectral analysis to cross-spectral analysis. For a joint process X_t and Y_t with an absolutely summable cross-covariance function, its cross-spectrum or cross-spectral density is given by

$$\begin{aligned} f_{XY}(\omega) &= \frac{1}{2\pi}\gamma_{XY}(e^{-i\omega}) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_{XY} e^{-i\omega k} \\ &= \frac{1}{2\pi}[c_{XY}(\omega) - iq_{XY}(\omega)] \end{aligned} \qquad (6.3)$$

where the real portion

$$c_{XY}(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_{XY}(\omega)\cos(\omega k)$$

is known as the *cospectrum*, and the imaginary portion

$$q_{XY}(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_{XY}(\omega)\sin(\omega k)$$

is known as the *quadrature spectrum*.

We can also write the cross-spectrum in the following polar form

$$f_{XY}(\omega) = A_{XY}(\omega) e^{i\phi_{XY}(\omega)} \qquad (6.4)$$

where

$$\begin{aligned} A_{XY}(\omega) &= |f_{XY}(\omega)| \\ &= \left[c_{XY}^2(\omega) + q_{XY}^2(\omega)\right]^{1/2} \end{aligned} \qquad (6.5)$$

and

$$\phi_{XY}(\omega) = \tan^{-1}\left[\frac{-q_{XY}(\omega)}{c_{XY}(\omega)}\right] \qquad (6.6)$$

The functions $A_{XY}(\omega)$ and $\phi_{XY}(\omega)$ are called the *cross-amplitude spectrum* and the *phase spectrum*, respectively. In addition, for helping interpretation, we often also consider two other useful functions, the *coherence* and the *gain function*. The coherence (or the squared coherency), $K^2_{XY}(\omega)$, is defined by

$$K^2_{XY}(\omega) = \frac{|f_{XY}(\omega)|^2}{f_X(\omega)f_Y(\omega)} \tag{6.7}$$

where $f_X(\omega)$ and $f_Y(\omega)$ are spectrums for X_t and Y_t, respectively. The gain function is defined as

$$G_{XY}(\omega) = \frac{|f_{XY}(\omega)|}{f_X(\omega)} = \frac{A_{XY}(\omega)}{f_X(\omega)} \tag{6.8}$$

which is the ratio of the cross-amplitude spectrum to the input spectrum.

The cross-amplitude spectrum measures covariance between X_t and Y_t processes at $\omega-$frequency. The coherence, like R^2 in regression, indicates the correlation between the $\omega-$frequency component of X_t and the $\omega-$frequency component of Y_t. Clearly, $0 \le K^2_{XY}(\omega) \le 1$. A value of $K^2_{XY}(\omega)$ close to 1 implies that $\omega-$frequency components of the two series are highly linearly related, and a value of $K^2_{XY}(\omega)$ close to 0 implies that they are only slightly linearly related or not linearly related. The gain function is simply the absolute value of the standard least squares regression coefficient of the $\omega-$frequency component of X_t. The phase spectrum, $\phi_{XY}(\omega)$, is a measure of the extent to which each frequency component of one series leads or lags the other. For example, in a simple causal model where there is no feedback relationship between X_t and Y_t, series X_t leads series Y_t at frequency ω if the phase $\phi_{XY}(\omega)$ is negative, i.e., $Y_t = \alpha X_{t-\tau} + e_t$; and series X_t lags series Y_t at frequency ω if the phase $\phi_{XY}(\omega)$ is positive, i.e., $X_t = \beta Y_{t-\tau} + e_t$. For a given $\phi_{XY}(\omega)$ at frequency ω, the actual time unit is given by $\phi_{XY}(\omega)/\omega$. Hence, the actual *lead time* from series X_t to series Y_t at frequency ω is equal to

$$\tau = -\frac{\phi_{XY}(\omega)}{\omega} \tag{6.9}$$

which is not necessarily an integer. A negative lead time in (6.9) indicates that series X_t lags series Y_t at frequency ω.

6.2 Estimation of the cross-spectrum

Given a bivariate series X_t and Y_t for $t = 1, 2, \ldots,$ and N, let $\hat{f}_X(\omega)$ and $\hat{f}_Y(\omega)$ be the smoothed spectrum estimates of $f_X(\omega)$ and $f_Y(\omega)$, respectively, and

$$\hat{\gamma}_{XY}(k) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{t=1}^{N-k}(X_t-\bar{X})(Y_{t+k}-\bar{Y}), k \ge 0 \\ \dfrac{1}{N}\displaystyle\sum_{t=1-k}^{N}(X_t-\bar{X})(Y_{t+k}-\bar{Y}), k < 0 \end{cases} \tag{6.10}$$

be the sample cross-covariances, we extend the smoothing method discussed in Section 5, and estimate the cross-spectrum by

$$\hat{f}_{XY}(\omega) = \frac{1}{2\pi}\sum_{k=-M_{XY}}^{M_{XY}} \lambda_{XY}(k)\hat{\gamma}_{XY}(k)e^{-i\omega k} \tag{6.11}$$

where M_{XY} and $\lambda_{XY}(k)$ are the corresponding truncation point and lag window for the smoothing. The same lag windows introduced in Section 5.1 can be used. Clearly,

$$\hat{c}_{XY}(\omega) = \frac{1}{2\pi}\sum_{k=-M_{XY}}^{M_{XY}} \lambda_{XY}(k)\hat{\gamma}_{XY}(\omega)\cos(\omega k)$$

and

$$\hat{q}_{XY}(\omega) = \frac{1}{2\pi}\sum_{k=-M_{XY}}^{M_{XY}} \lambda_{XY}(k)\hat{\gamma}_{XY}(\omega)\sin(\omega k)$$

Hence, we may estimate the cross-amplitude spectrum, the coherence, the gain function, and the phase spectrum as follows:

$$\begin{aligned}\hat{A}_{XY}(\omega) &= \left|\hat{f}_{XY}(\omega)\right| \\ &= \left[\hat{c}^2_{XY}(\omega) + \hat{q}^2_{XY}(\omega)\right]^{1/2}\end{aligned} \tag{6.12}$$

$$\hat{K}^2_{XY}(\omega) = \frac{|\hat{c}^2_{XY}(\omega) + \hat{q}^2_{XY}(\omega)|}{\hat{f}_X(\omega)\hat{f}_Y(\omega)} \tag{6.13}$$

$$\hat{G}_{XY}(\omega) = \frac{\hat{A}_{XY}(\omega)}{\hat{f}_X(\omega)} \tag{6.14}$$

and

$$\hat{\phi}_{XY}(\omega) = \tan^{-1}\left[\frac{-\hat{q}_{XY}(\omega)}{\hat{c}_{XY}(\omega)}\right] \tag{6.15}$$

The time lag can then be estimated by equation (6.9) with the estimate of the phase spectrum from (6.15).

The truncation point M_{XY} and the lag window $\lambda_{XY}(k)$ are chosen in a similar way to those in the univariate spectral analysis. In fact, we always begin the cross-spectral analysis with a careful univariate spectral analysis of each individual time series. The cross-spectral analysis is interesting and meaningful only when the univariate spectra are significant and contain enough power in one or both series. It should also be noted that the estimate of the phase spectrum is not reliable and meaningful when the coherence is small.

The sampling properties of the estimates of various cross-spectral functions are clearly related to the weighting function used in the smoothing. Because of limitations of space, we will not discuss them, and we instead refer interested readers to Peter Bloomfield (2000), David R. Brillinger (1975), and Maurice B. Priestley (1981).

Example 4

For illustration, we consider two monthly time series of spot prices of natural gas in Louisiana (X_t) and Oklahoma (Y_t) between January 1988 and October 1991 shown in Figure 35.6. It was known that the spot price for Louisiana natural gas was known on or before the first day of trading on the Oklahoma market. The datasets can be found in Wei (2006, p. 579).

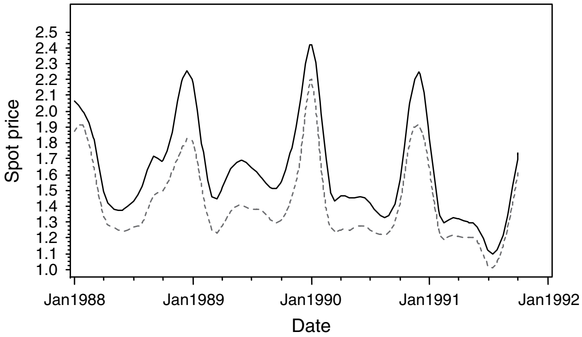


Figure 35.6 Spot prices of natural gas in Louisiana (solid line) and Oklahoma (dotted line) between January 1988 and October 1991

As shown in Figure 35.6, both series X_t and Y_t are stationary. We begin with a periodogram analysis for each individual series. The results indicate that each series contains a clear periodic component with a fundamental period of 12 months. We then use SAS to calculate the estimates of various cross-spectral functions with the Tukey window. The coherence and the phase spectrum are shown in Figures 35.7 and 35.8, respectively. The coherence remains substantially strong at all frequencies, implying that during the study period between January 1988 and October 1991 the correlation between the ω − frequency component of X_t and the ω − frequency component of Y_t are strong. The phase spectrum is nearly 0 at frequencies

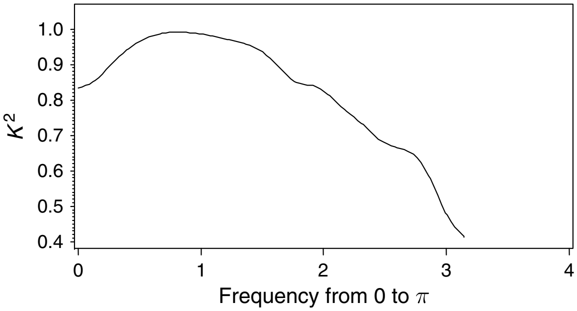


Figure 35.7 The coherence of Louisiana and Oklahoma spot prices of natural gas

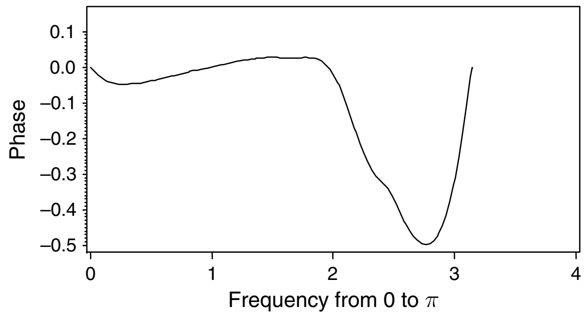


Figure 35.8 The phase spectrum of Louisiana and Oklahoma spot prices of natural gas

between 0 and $2\pi/3$ and negative at higher frequencies. It implies that during the same study period the two series are very much in alignment at low frequencies between 0 and $2\pi/3$, but series X_t leads series Y_t at high frequencies. In other words, during the study period the spot price of Louisiana natural gas leads the spot price of Okalahoma natural gas for a very short period of time.

7 Some mathematical detail

Let $\omega_k = 2\pi k/N, k = 0, 1, \ldots, [N/2]$ be the Fourier frequencies. We have used the following trigonometric identities in the derivation of Fourier series representation and the periodogram analysis of a series:

$$\sum_{t=1}^{N} \cos(\omega_k t) = \begin{cases} N, k = 0 \\ 0, k \neq 0 \end{cases} \tag{7.1}$$

$$\sum_{t=1}^{N} \sin(\omega_k t) = 0, \text{ all } k \tag{7.2}$$

$$\sum_{t=1}^{N} \cos(\omega_k t)\cos(\omega_j t)$$
$$= \begin{cases} N, \quad k = j = 0 \text{ or } N/2 (N \text{ even}) \\ N/2, k = j \neq 0 \text{ or } N/2 (N \text{ even}) \\ 0, k \neq j \end{cases} \tag{7.3}$$

$$\sum_{t=1}^{N} \sin(\omega_k t)\sin(\omega_j t)$$
$$= \begin{cases} 0, \quad k = j = 0 \text{ or } N/2 (N \text{ even}) \\ N/2, k = j \neq 0 \text{ or } N/2 (N \text{ even}) \\ 0, k \neq j \end{cases} \tag{7.4}$$

and

$$\sum_{t=1}^{N} \cos(\omega_k t)\sin(\omega_j t) = 0, \text{ for all } k \text{ and } j \tag{7.5}$$

To see these, we use the Euler relation

$$e^{i\omega_k} = \cos(\omega_k) + i\sin(\omega_k) \tag{7.6}$$

It follows that

$$\sin(\omega_k) = \frac{e^{i\omega_k} - e^{-i\omega_k}}{2i} \tag{7.7}$$

and

$$\cos(\omega_k) = \frac{e^{i\omega_k} + e^{-i\omega_k}}{2} \tag{7.8}$$

Now,

$$\begin{aligned} \sum_{t=1}^{N} e^{i\omega_k t} &= e^{i\omega_k}\left(\frac{1 - e^{i\omega_k N}}{1 - e^{i\omega_k}}\right) = e^{i\omega_k}\left(\frac{e^{i\omega_k N} - 1}{e^{i\omega_k} - 1}\right) \\ &= e^{i\omega_k}\left(\frac{e^{i\omega_k N/2}(e^{i\omega_k N/2} - e^{-i\omega_k N/2})/2i}{e^{i\omega_k /2}(e^{i\omega_k /2} - e^{-i\omega_k /2})/2i}\right) \\ &= e^{i\omega_k (N+1)/2}\frac{\sin(\omega_k N/2)}{\sin(\omega_k)} \\ &= \cos\left(\frac{\omega_k (N+1)}{2}\right)\frac{\sin(\omega_k N/2)}{\sin(\omega_k)} \\ &\quad + i\sin\left(\frac{\omega_k (N+1)}{2}\right) \\ &\quad \times \frac{\sin(\omega_k N/2)}{\sin(\omega_k)} \end{aligned} \tag{7.9}$$

However, from (7.6), we have

$$\sum_{t=1}^{N} e^{i\omega_k t} = \sum_{t=1}^{N} \cos(\omega_k t) + i\sum_{t=1}^{N} \sin(\omega_k t)$$

Equations (7.1) and (7.2) follow because

$$\sum_{t=1}^{N} \cos(\omega_k t) = \cos\left(\frac{\omega_k(N+1)}{2}\right)\frac{\sin(\omega_k N/2)}{\sin(\omega_k/2)}$$

$$= \begin{cases} N, k = 0 \\ 0, k \neq 0 \end{cases}$$

and

$$\sum_{t=1}^{N} \sin(\omega_k t) = \sin\left(\frac{\omega_k(N+1)}{2}\right)\frac{\sin(\omega_k N/2)}{\sin(\omega_k/2)}$$

$$= 0, k = 0, 1, \ldots, [N/2]$$

where we note that

$$\frac{\sin(\omega_k N/2)}{\sin(\omega_k/2)} = \frac{\sin(\pi k)}{\sin(\pi k/N)} = \begin{cases} N, k = 0 \\ 0, k \neq 0 \end{cases}$$

$$\cos\left(\frac{\omega_0(N+1)}{2}\right) = 1, \text{ and } \sin\left(\frac{\omega_0(N+1)}{2}\right) = 0$$

Equations (7.3), (7.4), and (7.5) follow immediately from (7.1), (7.2), and the following trigonometric identities

$$\cos(\omega_k)\cos(\omega_j) = \frac{1}{2}\{\cos(\omega_k + \omega_j) + \cos(\omega_k - \omega_j)\} \quad (7.10)$$

$$\sin(\omega_k)\sin(\omega_j) = \frac{1}{2}\{\cos(\omega_k - \omega_j) - \cos(\omega_k + \omega_j)\} \quad (7.11)$$

and

$$\cos(\omega_k)\sin(\omega_j) = \frac{1}{2}\{\sin(\omega_k + \omega_j) - \sin(\omega_k - \omega_j)\} \quad (7.12)$$

References

Bartlett, Maurice .S. (1950). Periodogram and continuous spectra. *Biometrika*, 37: 1–16.

Blackman, Ralph B. and Tukey, John W. (1958). *The Measurement of Power Spectra*. New York: Dover.

Bloomfield, Peter (2000). *Fourier Analysis of Time Series: An Introduction*, 2nd edn. New York: Wiley Interscience.

Box, George E.P. and Jenkins, Gwilym M. (1976). *Time Series Analysis, Forecasting and Control*, rev. edn. San Francisco: Holden Day.

Brillinger, Davis R. (1975). *Time Series, Data Analysis and Theory*. New York: Holt, Rinehart and Winston.

Brown, R. G. (1962). *Smoothing, Forecasting, and Prediction of Discrete Time Series*. Englewood Cliffs, NJ: Prentice-Hall.

Fisher, R. A. (1929). Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London, Series A*, 125: 54–59.

Hartley, Herman O. (1949). Tests of significance in harmonic analysis. *Biometrika*, 36: 194–201.

Insightful Corporation (2005). *S-Plus 7 for Windows*. Seattle, Washington.

Parzen, Emanuel (1961). Mathematical considerations in the estimation of spectra. *Technometric*, 3: 167–190.

Priestley, Maurice B. (1981). *Spectral Analysis and Time Series, Vols. I and II*. London: Academic Press.

SAS Institute (2003). *SAS for Windows, 9.1*. Cary, North Carolina.

Schuster, Arthur F. (1898). On lunar and solar periodicities of earthquakes. *Proceedings of the Royal Society of London*, 61: 455–465.

Schuster, Arthur F. (1906). On the periodicities of sunspots. *Philosophical Transactions of the Royal Society of London, Series A*, 61: 69–100.

SPSS (2006). *SPSS 14.0 for Windows*. Chicago, Illinois.

Tukey, John W. (1949). The sampling theory of power spectrum estimates. In Symposium on Applications of Autocorrelation Analysis to Physical Problems, NAVEXOS-P-735, 47-67, Office of US Naval Research.

Wei, William W. S. (2006). *Time Series Analysis – Univariate and Multivariate Methods*, 2nd edn. Boston: Addison-Wesley.

Whittle, Peter (1952). The simultaneous estimation of a time series harmonic components and covariance structure. *Trabajos, Estadist.*, 3: 43–57.

This page intentionally left blank

Chapter 36

Time-series techniques for repeated cross-section data

David Sanders and Hugh Ward

This chapter is concerned primarily with techniques that can be used to analyze aggregate data which describe different individuals over time. Such "repeated cross-section" data may be drawn from opinion surveys conducted by commercial polling agencies which interview a "new" random sample of respondents each month in order to ask them the same set of survey questions. Typically, the percentage of respondents who answer a given question in a particular way (e.g., the percentage answering "Conservative" in response to the question, "Which political party would you vote for if there were a general election tomorrow?") varies over time. This provides the researcher with an aggregate time series which, in principle, can be related empirically either to other attitudinal time series (e.g., responses to questions about consumer confidence) or to "objective" features of the economic and political environment, such as employment, interest rates and inflation.

However, a number of different techniques may be used for analyzing aggregate time-series data, and the choice between the techniques must depend ultimately on the kind of epistemological assumptions about the nature of "explanation" that the researcher is prepared to make in formulating a statistical model. The first four sections of this chapter review the main approaches to time-series modeling. The subsequent section applies the four techniques to the same data set and shows how the choice of technique can affect the statistical results obtained. The penultimate section briefly reviews the main strengths and weaknesses of the different techniques, concentrating particularly on the different epistemological assumptions that they make.

The four techniques for time-series analysis described below are all based upon the linear model. All allow for the complex multivariate analysis of interval-level data, making provision for the estimation of the effects exerted by a range of continuous and categorical explanatory variables. All can be used for analyzing historical data and for forecasting purposes.

1 The simple ordinary least squares (OLS) method

The form of the conventional regression (or OLS) model for cross-sectional data is well known:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e \qquad (1)$$

Reprinted from *Analyzing Social and Political Change*, Angela Dale and Richard Davies, eds, SAGE Publications, 1994.

where y is the response variable, $x_1, x_2, \ldots, x_k$ are explanatory variables, e is a normally distributed random error term, and the cases may be individuals or social aggregates.

With aggregate time-series data, the OLS model becomes

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + u_t \qquad (2)$$

where the t subscripts indicate that the y and x variables are measured over time; u_t defines the error term; and the cases are time points defined by the period for which data are available. The fundamental problem with OLS in this situation is that the u_t tend not to be independent[1], violating an important assumption upon which conventional methods of analysis depend.

The "serial correlation" in the error may often (but not always) be approximated by a "first-order autoregressive process" or AR(1) in which

$$u_t = \rho u_{t-1} + e_t \qquad (3)$$

where u_t and u_{t-1} are the (systematic) errors from an OLS time-series regression and e_t is an independently distributed error term. Such serially correlated error does not prejudice parameter estimation by OLS regression models but the standard errors of the coefficients will, in general, be underestimated and the R^2 overestimated. As a consequence, the risk of accepting a false hypothesis is increased.

On the other hand, serially correlated error may be a symptom of other misspecifications. These could include the omission of important explanatory variables which are correlated with variables in the model. They could also include failure to represent "feedback" dependence in which the level of response is dependent upon previous values of the response variable. If OLS methods are used in these circumstances, misleading results may be expected not only for standard errors but also for the parameter estimates.

2 Autoregressive (maximum likelihood) models

The autoregressive model offers the most immediate and obvious solution to the problem of serially correlated error: if serial correlation is distorting the standard errors of a given OLS model, why not attempt to specify the nature of the distorting "autoregressive" process and re-estimate the model taking account of that process? This should also result in more efficient estimation of the β parameters. A pragmatic approach is to fit an OLS model and then examine the pattern of intercorrelation among the estimated $u_t, u_{t-1}, u_{t-2}, \ldots$. If u_t correlates strongly only with u_{t-1}, then the relevant "error process" can be described as an AR(1); if u_t correlates strongly only with u_{t-1} and u_{t-2}, then the process is an AR(2); and so on.[2] The chosen error process is then incorporated into the estimation procedure; maximum likelihood (or asymptotically equivalent) methods for AR models are available in many standard software packages. If the specified autoregressive structure is correct, the estimation problems associated with OLS largely disappear: estimated standard errors are not deflated and, as a result, significance testing becomes a reliable exercise. This in turn means that the risks of wrongly accepting a false hypothesis (which spuriously links y_t to some x_t) are kept to a minimum. This said, autoregressive models have been criticized on two main grounds.

First, the ρ_k coefficients on the $u_t, \ldots, u_{t-n}$ terms are often difficult to interpret in

[1] Given that with most time-series data, y_{t-1} is generally a good predictor of y_t, it follows that u_{t-1} is likely to be a good predictor of u_t. In other words, u_t and u_{t-1} are likely to highly correlate.

[2] If u_t correlates only with, say, u_{t-3} and not with u_{t-1} or u_{t-2}, then this can be regarded as a "restricted AR(3)" model and estimated accordingly. See Pesaran and Pesaran (1987, pp. 69–70, 150–51).

substantive terms.[3] Second, models based purely on autoregressive techniques may simply be misspecified; in particular, they may have omitted important exogenous or endogenous variables which need to be included explicitly in the model rather than incorporated implicitly via the “catch-all” autoregressive structure. These criticisms are all the more potent in the case of “restricted” autoregressive models where u_t appears to be a function of, say, u_{t-5} but not of $u_{t-1}, \ldots, u_{t-4}$. Not only is such a result difficult to interpret substantively, but it also suggests that an additional exogenous variable (operating with a lag of around five time points) should be included explicitly in the model. In response, of course, the advocate of autoregressive techniques can argue that any autoregressive term is simply being employed instrumentally in order to obtain an accurate assessment of the magnitude of the effect of $x_{1t}, \ldots, x_{kt}$ on y_t, and that the question of the substantive meaning either of ρ_k or of u_{t-1} is therefore irrelevant.

3 The lagged endogenous variable OLS method

The defining feature of this technique is the inclusion of a term for y_{t-1} on the right-hand side of any equation which tries to predict y_t. This is in addition to the hypothesized effects of any exogenous variables which also need to be included. The basic form of the model is

$$y_t = \beta_0 + \alpha y_{t-1} + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + u_t \tag{4}$$

where y_t is the response (or endogenous) variable; y_{t-1} is the endogenous variable lagged by one time point; $x_1, \ldots, x_k$ are explanatory (exogenous) variables (which may exert lagged rather than simultaneous effects as shown in this example); and u_t is random error. Specifying the model in this way has two significant advantages. The first is that it frequently circumvents the problem of serially correlated error associated with simple OLS. The second advantage derives from the fact that y_{t-1} summarizes all the past effects of unmeasured variables (i.e., variables external to the model) on y_t (see Johnston, 1972, pp. 292–320). This means not only that the effects of measured variables $(x_{1t}, \ldots, x_{kt})$ on y_t can be estimated more accurately than would otherwise be the case, but also that the coefficient on $y_{t-1}(\alpha)$ represents the “discount rate”—the rate at which past influences on y_t decay. This latter feature of the model—that α is the discount rate—is particularly useful for specifying the rate of decay of “intervention effects”, such as the occurrence of particular political events. As shown below, for example, the Falklands War boosted government popularity in May 1983 by some 8.6%. The coefficient on the lagged dependent variable ($\alpha = 0.83$) enables us to infer that this effect decayed at a rate of about 0.83 per month thereafter. This implies that the “May boost” was worth $8.6 \times 0.83 = 7.2\%$ in June; $7.2 \times 0.86 = 5.9\%$ in July; and so on. It should be noted, however, that with this sort of model specification the effects of all measured exogenous variables are constrained to decay at the same rate. As discussed below, one significant advantage of Box-Jenkins models is that they permit the specification of a different decay rate for each exogenous variable.

Three other points need to be made about the lagged endogenous variable method. First, as with the simple OLS specification outlined earlier, the estimated coefficients may not be stable over time; i.e., they may take on radically different values if they are estimated over different subsets of the entire time series. A series of diagnostic tests for parameter stability are available (CUSUM, CUSUMSQ and recursive

[3] Although ρ_k can be taken to suggest the rate at which the past is discounted, it is still difficult to give substantive meaning to the expression $(u_t - \rho u_{t-1})$.

coefficient tests) and in general these should be applied systematically if either simple OLS or lagged endogenous variable methods are being used (see Brown et al., 1975). If a particular model fails these tests, then it is probably misspecified and requires either the inclusion of further exogenous variables or a respecification of the ways in which the existing $x_{1t}, \ldots, x_{kt}$ are hypothesized to affect y_t. Second, even with the inclusion of y_{t-1} in the equation, it is still possible that the error term from (4) will be subject to serial correlation. For example, one of the following could be the case:

$$u_t = \rho u_{t-1} + e_t \tag{5}$$

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_k u_{t-k} + e_t \tag{6}$$

$$u_t = \rho u_{t-2} + e_t \tag{7}$$

where (5) denotes a first-order autoregressive error process, (6) denotes a kth-order process, and (7) denotes a “restricted” second-order process. In any of these cases, as with simple OLS, some sort of correction for the error process is required. This can be resolved, of course, by incorporating an appropriate autoregressive error function in the model, although such a strategy carries with it all the limitations of the autoregressive model which were noted earlier.

Finally, it is worth observing that the lagged endogenous variable method described here represents a subspecies of the general to specific “Hendry” methodology followed by many UK econometricians (see, e.g., Hendry, 1983). This approach seeks to specify the short-run dynamics of time-series relationships by moving from a general model specification (which, in addition to y_{t-1}, includes all potential exogenous influences on y_t at all theoretically plausible lags) to a more limited, empirically determined specification which eliminates all non-significant exogenous terms. Thus, for example, a theoretical model which hypothesized that y_t was influenced by x_{1t} or x_{2t}, and in which it was assumed that any changes in x_{1t} or x_{2t} would take no longer than three periods to affect y_t, would initially be specified as

$$\begin{aligned} y_t = {} & \beta_{00} + \alpha y_{t-1} + \beta_{10} x_{1t} + \beta_{11} x_{1t-1} \\ & + \beta_{12} x_{1t-2} + \beta_{13} x_{1t-3} + \beta_{20} x_{2t} \\ & + \beta_{21} x_{2t-1} + \beta_{22} x_{2t-2} + \beta_{23} x_{2t-3} + u_t \end{aligned} \tag{8}$$

If it turned out empirically that x_{1t} influenced y_t with a lag of one time point while x_{2t} influenced y_t with a lag of two time points, then the final specification would be

$$y_t = \beta_{00} + \alpha_1 y_{t-1} + \beta_{11} x_{1t-1} + \beta_{22} x_{2t-2} + u_t \tag{9}$$

though, as noted before, this specification would have to be checked for both parameter stability and serially correlated error.

4 Box-Jenkins (ARIMA) methods

Box-Jenkins techniques differ from the regression-based techniques outlined above in two significant respects: (1) in their emphasis upon the need for time-series data to be systematically “pre-whitened” and the consequences this has for the way in which models are specified; and (2) in their facility for handling complex “intervention” specifications. We discuss each of these features in turn.

4.1 Pre-whitening and model specification

The basic data analytic principle underlying the Box-Jenkins methodology is that x_{t-k} helps to explain y_t in theoretical terms only if it explains variance in y_t over and above the extent to which y_t is explained by its own past values. The application of this principle in turn means that Box-Jenkins methods necessarily place great emphasis on the need to establish the precise nature of the “process” that is “self-generating” y_t. This is effected by the use of autocorrelation and partial autocorrelation functions which enable the analyst to determine what sort of “autoregressive” or “moving average” process (respectively, AR and MA processes, contributing to the ARIMA mnemonic) is generating y_t.[4] Once the self-generated sources of y_t have been specified, the analyst can then introduce an exogenous variable x_t into the model in the form of a "transfer function". The precise lag structure for the effects of x_t is determined by the use of a "cross-correlation function" which correlates the non-self-generated variation in y_t with the non-self-generated variation in x_t over a range of different lags and leads. If x_{t-k} (i.e., x_t at the specified lag or lags) yields a significant coefficient and produces a nontrivial reduction in the "residual mean square" of the transfer function (in other words, if x_{t-k} adds to the variation in y_t that is explained by the model as a whole), then it can be concluded that x_{t-k} does indeed exert an exogenous influence on y_t.

An important prerequisite of this modelling strategy is that all data which are to be used need to be "pre-whitened". This means that, before they are included in any transfer function analysis, all variables must be rendered mean and variance stationary: any trends in the component variables must be removed prior to analysis. This is normally effected by "differencing", where a first difference of y_t is defined as

$$\nabla y_t = y_t - y_{t-1} \tag{10}$$

and where a second difference of y_t is defined as

$$\nabla^2 y_t = \nabla y - \nabla y_{t-1} \tag{11}$$

Simple linear trends can usually be removed by first differencing; a decline-recovery (or rise-decline) trend by second differencing; and so on (see Appendix).

Given the assumption of pre-whitened data, the form of the Box-Jenkins model is relatively straightforward. Two matters complicate any presentation of it, however. First, Box-Jenkins techniques not only allow for the estimation of the direct effect of a change in x_t on y_t (analogous to the β coefficients in (4)), but also allow for the estimation of adjustment to steady state or "discount" parameters (analogous to the α coefficient in (4)) associated with those direct effects. One particularly attractive feature of the Box-Jenkins specification is that it permits, in effect, a different discount rate to be estimated for each exogenous variable. This can be a significant advantage over the lagged exogenous variable specification, which, as noted earlier, constrains the discount rate to be identical for all exogenous variables. What all of this means is that in the Box-Jenkins model there are potentially two parameters associated with each exogenous variable: a ω_k parameter, which measures the direct effect of x_{t-k} on y_t; and a δ_k parameter, which (in general) measures the rate at which the direct effect decays over time.

The second complicating aspect of the Box-Jenkins approach is the highly compressed nature of the notation, which often makes interpretation difficult for the nontechnical reader. For one thing, expositions of the method almost invariably employ the "backshift operator" B, where By_t means y_{t-1}; B^2y_t means y_{t-2}; and so on. For another, a general statement of the model would require rather more elaboration about the nature of AR and MA processes than can be developed here. We will therefore seek to illustrate the character of the model by the use of a hypothetical example. Consider a situation in which (a) a stationary endogenous variable y_t is influenced by two stationary exogenous variables x_{1t} and x_{2t}; (b) x_t affects y_t with a lag of one time point; and (c) x_{2t} affects y_t with a lag of three time points. The "compressed" statement of this model would be:

[4] For an accessible introduction to autocorrelation and partial autocorrelation functions, and indeed to Box-Jenkins techniques generally, see Liu (1990). The AR(1) model can be expressed as $y_t = \alpha y_{t-1} + u_t$, where u_t is a disturbance term and α is the parameter of the model. For an MA(1), the model can be written as $y_t = u_t - \rho u_{t-1}$, where u_t and u_{t-1} are disturbance terms and ρ is the parameter of the model.

$$y_t = B_0 + \frac{\omega_{11} B}{(1 - \delta_1 B)} x_{1t} + \frac{\omega_{23} B^3}{(1 - \delta_2 B)} x_{2t} + \mu_t \quad (12)$$

Without the use of the backshift operator, (12) becomes

$$\begin{aligned} y_t = {} & \beta_0^t + (\delta_1 + \delta_2) y_{t-1} - \delta_1 \delta_2 y_{t-2} \\ & + \omega_{11}(x_{1,t-1} - \delta_2 x_{1,t-2}) \\ & + \omega_{23}(x_{2,t-3} - \delta_1 x_{2,t-4}) + u'_t \end{aligned} \quad (13)$$

where

$$\beta'_0 = \beta_0 (1 - \delta_1 - \delta_2 + \delta_1 \delta_2)$$
$$u'_t = u_t - \delta_1 u_{t-1} - \delta_2 u_{t-1} + \delta_1 \delta_2 u_{t-2}$$

and where y_t, x_{1t} and x_{2t} are measured variables; β_0 is a constant; ω_{11} is the direct effect parameter for $x_{1,t-1}$; ω_{23} is the direct effect parameter for $x_{2,t-3}$; δ_1 and δ_2 are the decay parameters for ω_{11} and ω_{23} respectively; and u_t is a "white noise" (random) error component. With appropriate software (for example the BMDP-2T program), the estimation of the model described in (12), as well as rather more complex models, becomes a relatively straightforward exercise. Evaluating such models is largely a matter of examining the significance levels of the estimated parameters, checking that the transfer function model has a lower residual mean square value than the simple self-generating ARIMA model for y_t, and ensuring that u_t is indeed a white noise term in which no ARIMA process is evident. Provided that the model under analysis satisfies each of these conditions, it may be concluded that the specified exogenous variables do indeed affect y_t, and that the nature of their impact is defined by the appropriate ω_{ij} and δ_i parameters.

4.2 Complex intervention specifications in Box-Jenkins analysis

It was noted earlier that "intervention" effects can be estimated using the lagged endogenous variable method. Using this model, an intervention effect (such as the beginning of a war or the introduction of a new piece of legislation) is operationalized as a dummy variable I_t, which takes the value unity for the single time point when the intervention begins and zero otherwise. For any given y_t, this yields

$$y_t = \beta_0 + \alpha y_{t-1} + \beta_1 I_t + u_t \quad (14)$$

The estimated coefficient on I_t represents the increase or decrease in y_t associated with the intervention; the coefficient on y_{t-1} denotes the rate at which the intervention effect decays.

Box-Jenkins models allow for several additional ways of specifying such intervention effects. In the ensuing review we translate the models into a notation familiar to users of regression techniques.

The gradual-permanent model

In this model an exogenous variable intervenes at some time t^*, firstly, to change y_t immediately by some given amount, and secondly, to further change y_t through time in a way that approaches some upper limit. Compare the two models shown in Figures 36.1 and 36.2. The step function in Figure 36.1 is the standard dummy variable model of classical regression analysis, the dummy variable having the same (positive) impact on $E(y_t)$ (the expected value of y_t) for all values of t greater

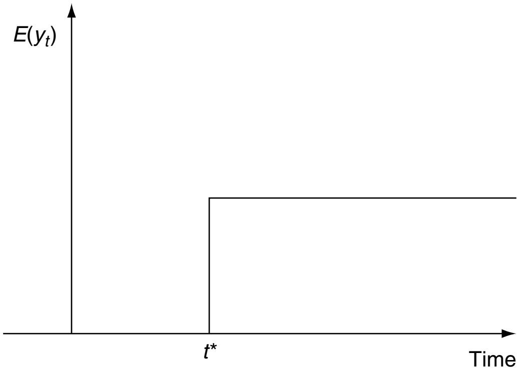


Figure 36.1 The dummy variable step-function model

than or equal to t^*. In contrast, the gradual-permanent model postulates a build-up of the intervention effect over time. Such a gradual build-up frequently seems more theoretically plausible then the "abrupt" effect modeled by a straightforward dummy: a new piece of race relations legislation, for example, may inhibit discrimination against ethnic minorities to only a limited degree initially, but the effects of the legislation may build gradually over time.

The simplest way of writing the gradual-permanent model is[5]

$$y_t = N_t \qquad \text{for } t < t^* \tag{15}$$

$$y_t = \omega \sum_{r=0}^{t-t^*} \delta^r + N_t \quad \text{for } t > t^*, 0 < \delta < 1 \tag{16}$$

where y_t is a measured variable; N_t is the ARIMA process self-generating y_t; ω is the initial intervention parameter; and δ is the "adjustment" parameter. As shown in Figure 36.2:

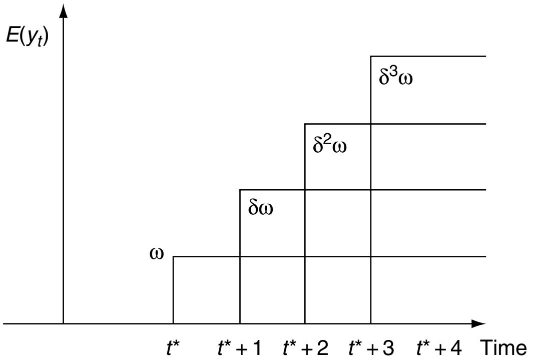


Figure 36.2 The gradual-permanent model

$$\text{at time } t^*, \quad E(y_t) = \omega\delta^0 = \omega$$

$$\text{at time } t^*+1, \quad E(y_t) = \omega(\delta^0 + \delta^1) = \omega + \omega\delta$$

$$\text{at time } t^*+2, \quad E(y_t) = \omega(\delta^0 + \delta^1 + \delta^2) = \omega + \omega\delta + \omega\delta^2$$

and so on. Given that δ must be less than unity, successive increments of δ^r become smaller and smaller. As δ approaches unity, the effect grows in an almost linear manner; as δ approaches zero, the growth in y_t tails off more and more rapidly. The value of δ, in short, is a parameter for the rate at which increments to y_t decay. Clearly, since the increments form a geometric progression, if $\delta < 1$ the series will converge on an upper limit of $\omega/(1-\delta)$. If ω is negative, the intervention has a reductive effect on y_t which gradually increases in magnitude.

The abrupt-effect/gradual-decline model

This specification is directly equivalent to the intervention effect associated with the lagged endogenous variable method summarized in (14). The model is most easily written as[6]

$$y_t = N_t \qquad \text{for } t < t^*$$
$$y_t = \omega\delta^{(t-t^*)} + N_t \qquad \text{for } t > t^*, 0 < \delta < 1$$

A specification of this sort, for a positive value of ω, is illustrated in Figure 36.3. The parameter

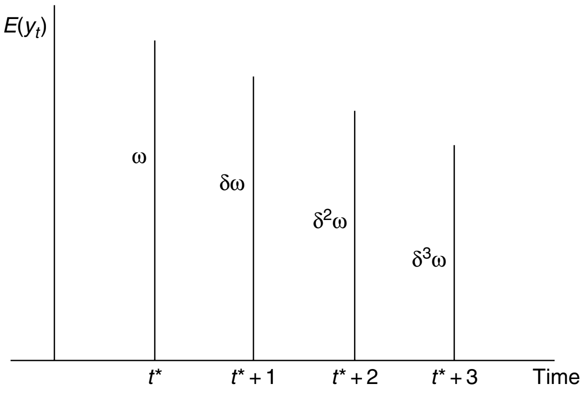


Figure 36.3 The abrupt-effect/gradual-decline model

[5] In the Box-Jenkins notation, $y_t = [\omega/(1-\delta B)]I_t + N_t$.

[6] In the Box-Jenkins notation, $y_t = [\omega/(1-\delta B)]\nabla I_t + N_t$.

δ measures the rate of decay in the initial ω effect; the smaller the value of δ, the faster the effect decays. As t increases, the effect of the intervention approaches zero.

The gradual-temporary model

In this specification the intervention has an immediate effect which builds up thereafter to some maximum value, and then decays gradually to zero. The simplest way of writing the model for the case in which the effect builds up for one period and then decays is[7]

$$\begin{aligned} y_t &= N_t && \text{for } t < t^* \\ y_t &= \omega_0 + N_t && \text{for } t = t^* \\ y_t &= \omega_0\delta^{(t-t^*)} + \omega_1\delta^{(t-t^*-1)} && \text{for } t > t^* \end{aligned} \qquad (17)$$

where $0 < \delta < 1$ and where ω_0 and ω_1 both have the same sign. As shown in Figure 36.4, the intervention model here is merely the sum of two abrupt-temporary models, with the first intervention commencing at t^* and the second at t^*+1. As long as ω_1 is greater than $\omega_0 - \delta\omega_0$, the effect of the overall "summed" intervention increases in the first post-intervention period and then gradually declines.[8]

As these examples show, Box-Jenkins techniques provide a variety of powerful and plausible ways of modeling intervention effects. However, it should be noted that, with the exception of the gradual-permanent model, similar intervention models can be specified using the somewhat simpler lagged endogenous variable technique.

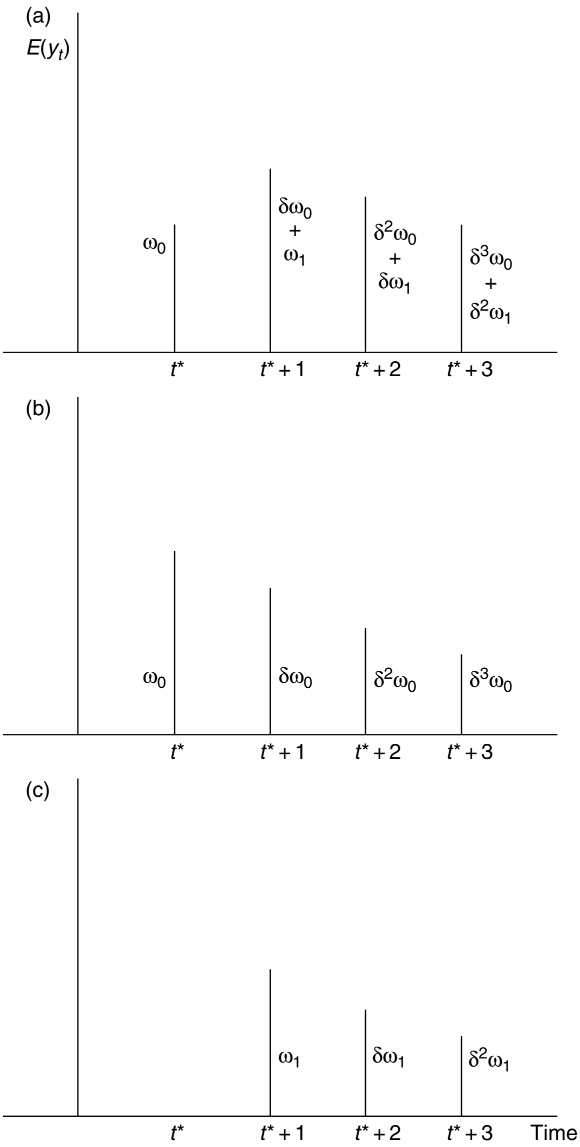


Figure 36.4 The gradual-temporary model: (a) total effect of ω_0 and ω_1 interventions (b) effect of ω_0 intervention (c) effect of ω_1 intervention

[7] In the Box-Jenkins notation, $y_t = [(\omega_0 + \omega_1)/(1-\delta B)]\nabla I_t + N_t$.

[8] This model, like the abrupt-effect/gradual-decline model, can be estimated using the lagged endogenous variable method. The model described in Figure 36.4, for example, could easily be estimated with the specification $y_t = \beta_0 + \alpha y_{t-1} + \beta_1 I_t + \beta_2 I_{t+1} + u_t$, where β_1 and β_2 are directly equivalent to ω_0 and ω_1 in (17); where α is analogous to δ in the same equation; where I_t is a dummy variable which takes on the value one for the period of the intervention and zero otherwise; and where u_t is a random error term.

5 An application of different time series to the same set of data: or how different assumptions can produce different conclusions

The particular example used here involves a problem that has interested political scientists throughout the postwar period: the question of the connections between the popularity of the incumbent government and the state of the domestic economy. The simple theoretical model that is investigated is shown in Figure 36.5. It hypothesizes that economic factors affect government support in two different, if complementary, ways. One set of ("evaluative") effects derives from the objective state of the economy as a whole and it is assumed that the better the overall performance of the economy, the more likely it is that the incumbent government will be rewarded with a high level of popular support. In the model presented here we measure the overall strength/weakness of the economy as an aggregate of three variables: inflation, unemployment and import prices.[9] The resultant "misery index", which in effect measures the weakness of the overall economy, is predicted to exert a negative effect on government popularity. The second set of economic effects is rather more "subjective" and "instrumental". They are concerned with the extent to which macroeconomic changes are perceived by voters as affecting their own narrow self-interests. If voters perceive that they are "doing very well" as a result of current policies, then they are more likely to lend their instrumental support to the party in power. The model assumes that "instrumental" judgements of this sort can best be measured at the aggregate level by the state of consumer confidence. If electors are optimistic about their own family's financial prospects, then they will be more likely to support the government whose policies produced that optimism in the first place; economic pessimism, in contrast, is likely to be associated with reduced governmental support.[10] Finally, it should be noted that the model also anticipates a political "intervention" effect: the Falklands War. Given that government popularity

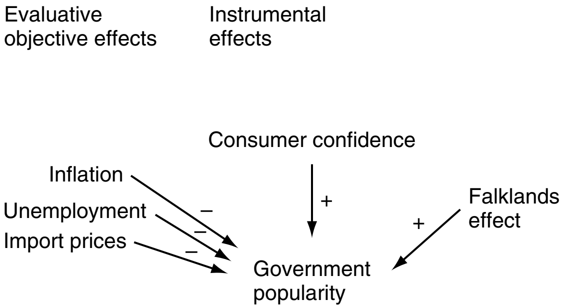


Figure 36.5 Hypothesized model of the main economic and political effects on UK government popularity, 1979–87. A plus sign denotes a predicted positive relationship, a minus sign a predicted negative relationship; the objective effects are combined to form a single misery index

[9] These variables were selected partly because of their close connections with the overall level of economic activity and partly because the first two at least have consistently received a great deal of media attention in the UK. All three were highly collinear (all bivariate correlations above $r = 0.9$ during the period analyzed here), which was why aggregation was considered necessary. The variables were aggregated by standardizing (to give each variable mean zero and unit standard deviation) and summing. The index accordingly gives equal weight to the three component variables.

[10] The consumer confidence measure employed here is based on the following monthly Gallup question which has been asked regularly since 1975: "Thinking about the financial position of your household over the next 12 months, do you expect to be: a lot better off; a little better off; about the same; a little worse off; a lot worse off?" The consumer confidence index is obtained by subtracting the percentage of respondents who think they will be worse off from the percentage who think they will be better off. For a recent examination of the connections between this index and government popularity in Britain, see Sanders (1991b).

increased dramatically in the early months of the war (particularly in May and June of 1982), it makes sense to seek to assess how far the war might have contributed to the Conservatives' election victory in June 1983.

Table 36.1 reports the results derived from estimating the same model using each of the four techniques reviewed above. The data are monthly time series and cover the period June 1979 to June 1987, Mrs Thatcher's first two terms of office as prime minister. The models all assume that while changes in consumer confidence have a near instantaneous effect on government popularity, changes in the objective state of the economy only work through to popularity after a two-month lag.

The simple OLS model (18a) at first sight appears to provide some support for the hypothesized model shown in Figure 36.5. Although R^2 is fairly low (0.54), the parameters for consumer confidence and the misery index are both significant and in the predicted direction. And although the Falklands War coefficients are not all significant, the war certainly appears to have

Table 36.1 Parameter estimates for models of UK government popularity, June 1979 to June 1987 (standard error in parentheses)

Independent variable	*(18a)*	*(18b)*	*(19)*	*(20)*	*(21)*	
	Simple OLS model	*Simple OLS model, fewer Falklands variables*	*Autoregressive model*	*Lagged endogenous variable model*	*Box-Jenkins model*	
					ω parameter	*δ parameter*
Popularity$_{t-1}$/AR$_{(t)}$ parameter			0.91	0.83	0.89	
			(0.04)	(0.04)	(0.05)	
Consumer confidence	0.49	0.50	0.15	0.12	0.13	0.66
	(0.05)	(0.05)	(0.05)	(0.03)	(0.04)	(0.19)
Misery index$_{t-2}$	−2.06	−2.10	−0.30	−0.38	0.11	
	(0.38)	(0.39)	(1.18)	(0.20)	(0.24)	
Falklands-May	5.81	5.68	4.37	8.63	7.98	0.93
	(3.75)	(3.84)	(1.61)	(1.71)	(1.81)	(0.08)
Falklands-June	8.10	7.95	5.17	5.54	4.32	0.51
	(3.75)	(3.84)	(1.62)	(1.71)	(1.76)	(0.41)
Falklands-July	7.14					
	(3.75)					
Falklands-August	5.81					
	(3.75)					
Constant	39.73	39.89	37.85	6.94	38.62	
	(0.48)	(0.48)	(2.33)	(1.73)	(1.99)	
Residual square	12.89	13.77	3.24	2.68	3.05	
N	95	95	95	95	97	
Durbin-Watson	0.45	0.44	1.95	2.06		
R^2	0.54	0.51	0.88	0.90		
(adjusted R^2)	(0.51)	(0.48)	(0.88)	(0.90)		
Estimated contribution of Falklands boost to government popularity in June 1983	0	0	0	1.35	3.10	

boosted government popularity by around 3% in the early summer of 1982. Dropping the nonsignificant Falklands dummy variables for July and August has only a very marginal effect on the model (18b). However, the very low values of the Durbin-Watson statistics suggests that the models suffer from a serious first-order serial correlation problem, a conclusion that is confirmed by further diagnostic tests which we do not report here.[11]

The three other models, (19), (20) and (21), in their own different ways correct for the problems of serially correlated error.[12] These models have much lower residual mean square values (which is another way of saying that the R^2 values, where they are calculated, are much higher) than the simple OLS models. This is not surprising because (19)–(21) all include either y_{t-1} or some transformed version of it, in the form of u_{t-1}, in the estimation procedure. There are other consistent patterns in the results for these three models. The coefficient of the consumer confidence variable remains significant, although it is much smaller (between 0.12 and 0.15) than in the OLS models (where it is 0.49 or 0.50). Each of the three models also produces a significant coefficient for the Falklands-May effect, again in contrast to the OLS model which yields a nonsignificant coefficient for this variable.

Yet in spite of these similarities among the models summarized in (19)–(21), the results reported also indicate some important differences. The most notable of these is the fact that the lagged endogenous variable method produces a significant (and, as predicted, negative) coefficient for the misery index ($\beta = -0.38$), whereas the autoregressive and Box-Jenkins methods both yield nonsignificant coefficients (and in the latter case the wrong sign: $\beta = 0.11$). In substantive terms, there is clearly a problem here. If we are prepared to believe the results of the lagged endogenous variable model, then we would conclude that the "objective" state of the economy does exert a direct influence upon voters' support for the government; yet if we believe the results of either the autoregressive or Box-Jenkins models then we would conclude that there was no such role for objective economic factors.

A similar problem emerges when we try to establish what each of these models implies about the impact of the Falklands War on the outcome of the 1983 general election. The OLS model (18a) implies that the war had no effect on the subsequent election whatsoever. It boosted popularity substantially in June 1982, but its effects were already statistically nonsignificant by July of that year (see the nonsignificant coefficients for the Falklands-July and Falklands-August variables in (18a). Similarly, the autoregressive model (19) does not include any mechanism whereby the significant May and June 1982 effects may have a continuing impact on the response variable. Both models (20) and (21), however, suggest that the Falklands effect followed the gradual-temporary intervention model described earlier. Popularity was boosted in May (according to (20) by 8.6%, and according to (21) by 8.0%); was boosted further in June (in (20) by 5.5% and in (21) by 4.3%); and subsequently "decayed" gradually. The rate of decay, however, varies according to the different models. In the lagged endogenous variable model, the decay rate is given by the coefficient on y_{t-1}. This implies that the May boost of 8.6% was worth $8.6 \times 0.83 = 7.2\%$ in June; $8.6 \times (0.83)^2 = 5.9\%$ in July; $8.6 \times (0.83)^3 = 4.9\%$ in August; and so on. The June boost of 5.5% was worth $5.5 \times 0.83 = 4.6\%$ in July; $5.5 \times (0.83)^2 = 3.8\%$ in August; and so on. Combining these two sets of effects

[11] The standard tests for serially correlated error are available on many econometric packages. See, for example, Pesaran and Pesaran's Datafit program (1987).

[12] The lagged endogenous variable model shown in (20) was checked systematically for serially correlated error; no evidence of serial correlation was found.

together, the Falklands War was still worth some 1.3% to the government in June 1983.[13] Given the relative inaccuracy of opinion poll data even when government popularity is measured, as it is here, by the "poll of polls" – this result casts considerable doubt on claims (e.g., Norpoth, 1987; Clarke et al., 1990) that the domestic political effects of the Falklands campaign played a decisive role in the Conservatives' election victory in 1983. On the contrary, the importance of the consumer confidence and misery variables in the lagged endogenous variable model support the argument that the 1983 election outcome was the result primarily of economic factors; that the government had secured sufficient economic recovery by the summer of 1983 to ensure its re-election (see Sanders et al., 1987 and Sanders, 1991a).

Yet if (20) suggests that the medium-term effects of the Falklands War on government popularity were negligible, the results reported for the Box-Jenkins models (21) imply that those effects were a little more substantial. In the Box-Jenkins model, the δ parameters explicitly estimate the rate of decay in their respective ω parameters. The δ parameter for the Falklands-June variable is not significant ($t = 1.23$) but the Falklands-May effect was still worth 3.1% in June 1983.[14] The clear substantive implication of this result is that although a Falklands effect of 3% would probably not have been decisive in the circumstances of 1983, it would nonetheless have made an important contribution to the size of the Conservative election victory in June of that year. It is more than twice the effect estimated by the lagged endogenous variable model.

Where does this leave us? Was there a measurable (if modest) Falklands factor at the time of the 1983 election? Should we conclude that the objective state of the economy (as measured by the misery index) had no direct effect on the electorate's support for the government during the first two Thatcher terms? Unfortunately, as the foregoing discussion indicates, these questions cannot be answered independently of the statistical techniques that are employed to investigate them. We can certainly dispense with the conclusions suggested by the simple OLS model, but there is no easy way of resolving the disparities between the autoregressive, lagged endogenous variable and Box-Jenkins methods. The only thing that can be said definitely is that in this particular case—and most likely in others as well—the different techniques produce different statistical results and, by implication, different substantive conclusions. How, then, can we decide between the different techniques?

6 Which techniques? Assessing relative strengths and weaknesses

Several of the strengths and weaknesses of the different techniques have already been mentioned. Here we summarize and qualify them.

The simple OLS method has the enormous advantage of being easily understood; yet, as noted above, its frequent contamination by serially correlated error often makes it problematic for time-series analysis. This said, from an epistemological point of view, the way in which the simple OLS technique tests a given model does correspond most obviously to what most social scientists would regard as "testing a causal explanation". Without wishing to minimize the enormous difficulties associated with the concept of "explanation", we would argue that a causal explanation of a particular phenomenon or set of phenomena consists in

[13] As there were 13 months from May 1982 to June 1983, and 12 months from June 1982 to June 1983, the combined effect by June 1983 is given by $(8.6 \times (0.83)^{13}) + (5.5 \times (0.83)^{12}) = 1.3\%$.

[14] Since the Falklands-May δ parameter estimate is 0.93, the May-boost effect by June 1983 is $7.98 \times (0.93)^{13} = 3.1$. Using the nonsignificant estimated $\delta = 0.5$ to denote the decay rate of the June boost gives a negligible additional effect by June 1983 $(4.32 \times (0.51)^{12} = 0.00)$.

the specification of the minimum nontautological set of antecedent necessary and/or sufficient conditions required for its (their) occurrence. The simple OLS model, by placing the response variable on the left-hand side, allows the researcher to assess how far a knowledge of the explanatory variables at time t, time $t-1$ and so on, permits accurate predictions to be made about the response variable at time t. If extremely accurate predictions can be made, then it can be concluded that (at the specified level of abstraction) the minimum nontautological set of antecedent necessary and sufficient conditions required for the occurrence of the response variable has indeed been identified; in other words that an explanation of the response variable has been tested and found to be consistent with the available empirical evidence.

Given the epistemological position of the simple OLS approach, it comes as something of a surprise to discover that many practitioners of time-series analysis reject the use of simple OLS methods out of hand. Where there are strong linear trends in both explanatory and response variables this rejection is entirely justified; the coincidence of the trends usually suggests that some third, unmeasured variable (or set of variables) is operating to produce a spurious correlation between y_t and x_t.

The solution to this problem adopted by the autoregression, lagged endogenous variable and Box-Jenkins techniques (for shorthand purposes we will refer to these collectively as AR-LEV-BJ techniques) is to take explicit account of the extent to which y_t can be predicted by its own values when any attempt is being made to estimate the effects of x_{t-k} on y_t. In the lagged endogenous variable case, the effects of any unmeasured variables are constrained to operate through y_{t-1}; in the autoregressive and Box-Jenkins models the unmeasured variable effects are constrained to operate through whatever autoregressive or ARIMA process appears to determine y_t. By controlling for the unmeasured influences on y_t in this way, the autoregressive, lagged endogenous variable and Box-Jenkins methods not only pre-empt the problem of serially correlated error, but also provide for a more accurate estimation of the effect on y_t that results from a unit change in some x_{t-k}.

As is so often the case, however, resolving one problem serves only to raise another; in this case, an epistemological one. By including the past history of y_t into the estimation procedure, the AR, LEV and BJ methods in effect introduce a version of y_{t-k} on to the right-hand side of the equation for y_t. Yet, if we go back to our definition of a causal explanation, we see that it requires the specification of the "minimum set of nontautological antecedent . . . conditions" necessary for the occurrence of the phenomenon in question. Since y_{t-1} is certainly not defined independently of y_t, it appears that the AR, LEV and BJ specifications build a very powerful tautological component into the explanations of y_t that they imply. Moreover, since tautological "explanations" are not explanations at all—a variable clearly cannot explain itself—it would seem to follow that attempts to explain y_t based upon AR, LEV or BJ model-building procedures can never provide "real" explanations at all; they are merely refined vehicles for specifying the consequences for y_t of a unit change in x_t. They never allow the analyst to conclude that "movements in y_t can be (nontautologically) explained by movements in x_t and the data are consistent with the proposition that y_t is caused by x_t". Since this, in our view, is one of the main goals of empirical analysis, we believe it constitutes a serious limitation on the AR, LEV and BJ methods.

Not surprisingly, supporters of AR, LEV and BJ techniques respond strongly to these criticisms. They point out that any terms for y_{t-k} can easily be moved across to the left-hand side of the equation so that the right-hand side effectively contains expressions that are nontautologically related to y_t. In the case of the LEV model, for example, moving y_{t-1} from the right-hand side to the left-hand side of the equation is equivalent to using the change in y_t as the response variable. All that is being done, in short, is to shift the nature of that which is to be explained from the level of y_t to the change in y_t. If we know the "start" value for y_t, we can easily get back to predicting its subsequent levels. So far, so good.

Yet the epistemological costs of this notational shuffling are more serious than its protagonists imply. This can be seen, firstly, by reference to R^2. With simple OLS (assuming parameter significance and stability), and an absence of serially correlated error) R^2 is a singularly useful statistic: it reveals how well y_t can be predicted purely from movements in (nontautological) exogenous variables. However, with the AR and LEV models (as their advocates would readily admit), R^2 is highly misleading as a guide to the explanatory power of the nontautological influences on y_t because it is calculated using an estimation procedure that explicitly incorporates the past history of y_t. Since the AR, LEV and BJ specifications seek to explain only the non-self-generated variation in y_t, what is really required is an R^2 equivalent that measures the extent to which the exogenous variables in a particular model can indeed predict y_t. The statistic that is usually employed in this context is the residual mean square (RMS). As noted earlier, if the addition of a particular x_{t-k} yields a lower RMS value than that obtained by knowing only y_t's past history, then it is inferred that x_{t-k} does affect y_t. What RMS tests of this sort do not reveal, however, is how accurately the non-self-generated variation in y_t can be predicted purely from a knowledge of the exogenous influences upon it. Yet, as discussed above, the only sense in which AR, LEV and BJ techniques explain anything in conventionally understood causal terms is that they can account for the non-self-generated variation in y_t. Curiously, the summary statistics usually assocated with these techniques fail to give any clear indication as to the extent to which this objective is in fact achieved. With AR, LEV and BJ methods, in short, not only is the explicandum (the non-self-generated variation in y_t) a much-reduced version of the original phenomenon of interest, but there is also a failure adequately to assess the extent to which that reduced explicandum is indeed nontautologically explained.

A second epistemological cost associated with AR, LEV and BJ methods also derives from their emphasis on the need to take full account of the self-generated variation in y_t. The notion that x_{t-k} only affects y_t to the extent that it explains variation in y_t not explained by y_t's own past history certainly accords with the principle of "Granger causality".[15] Unfortunately, it also engenders a serious risk of underestimating the explanatory importance of x_{t-k}. In any given time-series model, it is entirely possible that the self-generated variation in y_t is also capable of being explained by some x_{t-k}. It is highly unlikely, however, that x_{t-k} will predict y_t as well as y_t's own past values will predict y_t. This, in turn, implies that x_{t-k} may appear to exert no influence on y_t simply because it explains the same variation in y_t that is explained by some function in y_{t-k}. In situations such as this, AR, LEV and BJ models are biased towards the underestimation of exogenous effects.

All this discussion of self-generated variation, however, does little to resolve one of the main substantive problems posed in the previous section. If, in a specific case, serially correlated error prejudices a simple OLS model,

[15] For an introduction to Granger causality, see Freeman (1983). Granger's notion of causality can be summarized as follows: x_t can be considered to Granger cause y_t if (a) y_t is influenced both by x_t and by lagged values of x_t but x_t is not influenced by lagged values of y_t, and (b) if x_t explains variation in y_t after all selfgenerated variations in y_t have been taken into account or eliminated.

which of the AR, LEV and BJ class of models should we employ in order to analyze repeated aggregate cross-section data? (The choice of models clearly matters. For example, as we saw earlier, the LEV model found evidence of a significant "misery" effect on popularity, whereas the AR and BJ methods suggested no such effect.) There is, sadly, no easy or general answer to the question. The researcher must decide which is the more appropriate in the light of her or his particular theoretical concerns. This is not to imply that researchers can simply select the technique which best seems to provide empirical support for their preconceived theoretical suppositions. It does mean, however, that some attempt needs to be made to link the assumptions of the modeling technique to the kind of model of human behavior that the researcher is seeking to test.

In the context of the government popularity functions used here, for example, we would argue that the entire exercise only makes sense if it is possible at some stage to translate the parameters of a given model into the decision calculus of the individual elector. In our view, this requirement renders the LEV method the most appropriate of the class of AR, LEV and BJ techniques for analyzing government popularity data. In contrast with the AR and BJ techniques, all of the terms specified in the LEV model are directly measured, so that it can be plausibly assumed that the typical elector is in some sense aware of them. The inclusion of the lagged endogenous variable itself can be interpreted as denoting the elector's predisposition to support the government; the exogenous variables denote, obviously, the hypothetical economic and political influences on government popularity.

With the AR and BJ techniques, in contrast, the translation is much more difficult. Although the same sort of interpretation can be made of the exogenous variables as in the LEV case, the substantive meaning of u_{t-1} in the AR model and of the ARIMA process that is self-generating y_t in the BJ model—both by definition phenomena that are not directly observable—is generally far from clear. In these circumstances, it is often difficult to envisage how the coefficients on some of the terms in AR and BJ models translate into what might conceivably go on inside electors' heads.

It is primarily for this reason that we would conclude that the lagged endogenous variable technique is probably the most appropriate for analyzing the sort of data we have described in this chapter. This, in turn, leads us to conclude that, of the various results presented in Table 36.1, the findings reported in (20) are probably the most useful for evaluating the theoretical model that was proposed in Figure 36.5. This suggests (a) that both consumer confidence and the objective state of the economy (which may themselves be interrelated)[16] exert direct effects on the level of electoral support for the government; and (b) that the Falklands War was worth just over 1% to the government's popularity by the time of the 1983 election. This is not to imply, however, that the LEV technique is always the best vehicle for handling time-series data. Where the nature of the substantive problem under investigation means that it is unnecessary to translate the parameters of the statistical model into some kind of individual-level decision calculus, it may well be more appropriate to obtain as accurate a definition of the error process as possible; and in these circumstances, AR or BJ methods would probably be more suitable than the LEV technique as it has been outlined here.

7 Conclusion

In this chapter we have reviewed four different—though related—techniques for time-series analysis. We have also attempted to articulate our doubts, not about the statistical

[16] For a discussion of these connections, see Sanders et al. (1987).

soundness of AR, LEV and BJ techniques, but about their epistemological implications; about the limitations on their ability to evaluate "explanations" as they are conventionally understood. We certainly do not claim to have provided a definitive analysis of the epistemological difficulties encountered with these techniques, merely to have articulated some genuine sources of concern which, in our view, all practitioners of time-series analysis should at least consider.

Time-series analysis with political and social data is necessarily a highly judgemental process. Apart from the continuing need to avoid models that exhibit serially correlated error, there are few hard and fast rules that must be followed in all circumstances. Indeed, any method that seeks to impose a strict set of rules to be followed will tend to founder on the need for constant interplay between the analyst's theoretical ideas and the way in which they use particular statistical techniques. Time-series analysis with political and social data is not the hard "science" of the econometricians. It involves the evaluation of causal propositions by reference to concepts that are imperfectly measured and techniques that are rarely altogether appropriate for the task. It is, in essence, art with numbers.

Appendix: effects of differencing on two hypothetical time series

Detrending by differencing can lead to substantively spurious conclusions if the trends in endogenous and exogenous variables are causally related. A familiar paradox is displayed in Figures 36.6 to 36.8. The (hypothetical) y_t and x_t variables are clearly negatively related over the long term (each reaches its maximum/minimum at the same point). Yet if both y_t and x_t are differenced to render each series stationary (as in Figure 36.8), the transformed, second-differenced variables appear to be positively related, even though common sense suggests that such an inference is inappropriate.

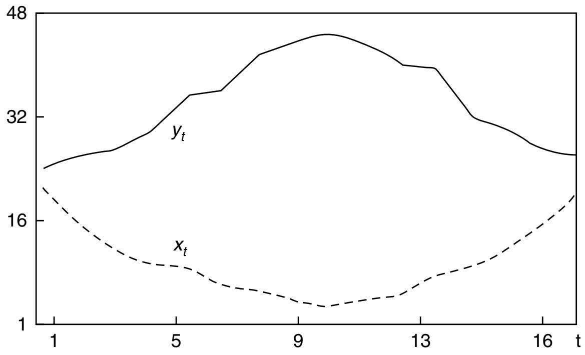


Figure 36.6 Hypothetical y_t and x_t over time

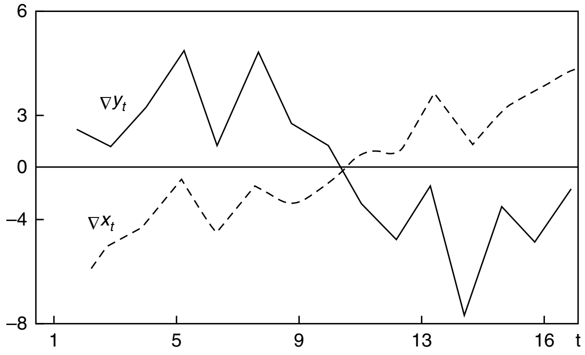


Figure 36.7 Changes in y_t and x_t (first differences) over time

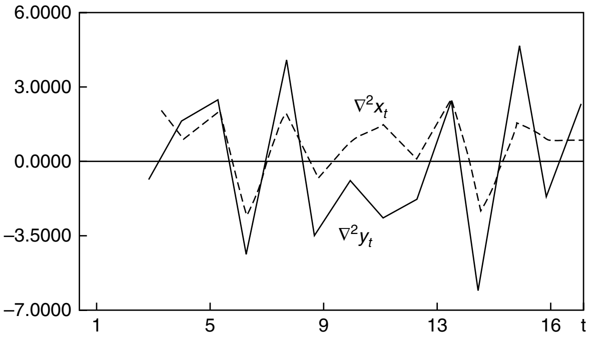


Figure 36.8 Changes in ∇y_t and ∇x_t (second differences) over time

In circumstances such as this, an insistence that all variables must be rendered stationary can produce misleading conclusions. At worst—in situations where there is a considerable amount of measurement error in both y_t and x_t and where the measured variables therefore only track the broad trends in the phenomena under investigation—the removal of trends through differencing can lead to a statistical analysis based almost exclusively on the correlation of measurement error.

References

Brown, R.L., Durbin, J. and Evans, J.M. (1975). Techniques for testing the constancy of regression over time. *Journal of the Royal Statistical Society, Series B*, 37: 149–92.

Clarke, H., Mishler, W. and Whitely, P. (1990). Recapturing the Falklands: Models of Conservative popularity, 1979–83. *British Journal of Political Science*, 20: 63–82.

Freeman, J.R. (1983). Granger causality and the time series analysis of political relationships. *American Journal of political Science*, 27: 327–57.

Hendry, D.F. (1983). Econometric modelling: the “consumption function in retrospect”. *Scottish Journal of Political Economy*, 30: 193–229.

Johnston, J. (1972). *Econometric Methods*, 2nd edn. New York: McGraw-Hill.

Liu, L.M. (1990). Box-Jenkins time series analysis. In *BMDP Statistical Software Manual*. Oxford: University of California Press.

Norpoth, H. (1987). Guns and butter and government popularity in Britain’. *American Political Science Review*, 81: 949–59.

Pesaran, M.H. and Pesaran, B. (1987). *Datafit: An Interactive Econometric Software Package*. London: Wiley.

Sanders, D. (1991a). Voting behaviour in Britain: lessons of the 1980s. *Contemporary Record*, 4: 2–6.

Sanders, D. (1991b). Government popularity and the next general election. *Political Quarterly*, 62: 235–61.

Sanders, D., Ward, H. and Marsh, D. (1987). Government popularity and the Falklands War: a reassessment. *British Journal of Political Science*, 17: 281–313.

This page intentionally left blank

Chapter 37

Differential equation models for longitudinal data

Steven M. Boker

One reason that longitudinal data are collected is to understand dynamic processes that might give rise to time-dependent relationships in these measured variables. When intensive longitudinal designs have been employed, consisting of more than thirty or so occasions of measurement, continuous time differential equations may be good candidates for empirically testing hypotheses about dynamic processes. This chapter introduces the reasoning behind the use of continuous time differential equations models, explores a few of the simple popular forms of differential equations that have been used in the behavioral sciences, and provides an introduction to one of the methods that is currently in use for fitting differential equations models to intensive longitudinal data.

Individuals change over time. This change may be minute to minute fluctuation in variables like mood or anxiety, longer-term change in variables like attitudes, or very long-term developmental change in variables like measures of cognitive abilities. If an individual is changing in such a way that his current state influences his future state, then one may find that differential equations compactly describe hypotheses about the process by which this time-dependence can lead to a variety of individual trajectories of change. That is to say, we do not expect that everyone's pattern of change will look alike. However, the same lawful internal process may lead to what appear to be qualitatively different patterns of change for different individuals.

There are several ways that observed individual differences in intraindividual change trajectories may occur. Let us consider three of them. First, there may be essentially random exogenous influences that may also influence the future state, so the future state may only be partially dependent on the current state. Second, there may be quantitative individual differences in the way that the current state leads to the future state for each individual. Third, an individual may adapt the way that his current state leads to his future state in response to some change in the environment. Differential equations modeling allows one to succinctly formalize each of these ways in which interindividual differences in process may lead to differences in intraindividual change.

Intraindividual change may be referenced as being in relation to an *equilibrium set*. That is, the internal process leading to intraindividual change is somehow organized around the equilibrium set. For instance, a simple equilibrium set could be a single personal goal. One might observe behavior of an individual that regulated itself so as to bring the individual closer to this personal goal. Another equilibrium set might be a homeostatic set point. One might measure a behavioral variable that fluctuated around the set point, while not leaving its immediate neighborhood of values. But an equilibrium set does not need to be a single value; an equilibrium set may be a cycle, such as the circadian wake–sleep cycle. Or, an equilibrium set might itself be undergoing developmental change such as when crawling turns to walking in toddlers.

Differential equations models allow one to formalize the relationship between observed short-term patterns of individual change from one observation to the next and the overall pattern of changes that could have been observed if the individual had been observed starting in any particular state. That is to say, differential equations models infer the overall organization of the intraindividual change with respect to an equilibrium set from the observed sample of short-term intraindividual changes.

In order to make the preceding ideas more concrete, let us examine three simple linear differential equations. A continuous time differential equation is a model for instantaneous change. The term instantaneous change as it is used here does not imply that a person changes some discrete amount, say 10 points on an abilities measure, from one moment to the next. Instead, it is saying that there is a relationship between the value of change and the interval of time over which the change happens, and that this ratio exits for every moment in time during the interval. This ratio is expressed as a derivative with respect to time. For instance, the first derivative with respect to time of a variable x at time t can be expressed as $dx(t)/dt$ which we will write in shorthand as $\dot{x}(t)$. This simply says that at some chosen instant of time t the slope of x exists and we have an estimate of it, $\dot{x}(t)$. Similarly, we can talk about the change in that slope. At the chosen time t, the change in the first derivative would be $d(dx(t)/dt)/dt$ which is shortened to $d^2x(t)/dt^2$ and which we will write in shorthand as $\ddot{x}(t)$. Thus, at a chosen time t there exists a curvature for the variable x and we have an estimate of it, $\ddot{x}$.

Consider a single outcome variable x that is a linear function of time, so that

$$x(t) = b_0 + b_1 t \tag{1}$$

where b_0 is the intercept and b_1 is the slope. Be sure to note that b_1 has units associated with it: change in x per unit change in time. For instance, if x was your car and you were traveling at a constant velocity while being photographed from the air, the value for b_1 might be in units of miles per hour, or meters per minute. Without the units for b_1, it does not have a meaning that can be interpreted from one experiment to the next. This is an important point that relates back to the definition of the first derivative in the previous paragraph: in experimental data, derivatives always have units associated with them (such as miles per hour, meters per second). Taking the differential of equation 1 with respect to time, we find that

$$\dot{x}(t) = b_1 \tag{2}$$

This is just as we would expect, the slope is b_1 no matter what time t we select. So, here is our first example differential equation. We are predicting the change in x and it is constant for any time t.

But where is b_0 in equation 2? The intercept b_0 is the value of x when $t = 0$. We are no longer predicting x, so there is no need for b_0; it drops out during the differentiation with respect to time. This is one major difference between using a differential equation and using the integral form as shown in equation 1, which explicitly predicts a value for the variable x. Differential equations do not specify what happens at time $t = 0$. This can be a substantial benefit when one is not interested in how x changes with respect to time $t = 0$, but rather is interested in how x might change given any particular value of x at an arbitrarily assigned value of $t = 0$.

In integral form models for change, it is important to be able to assign a time $t = 0$ and that this value for time is the same for everyone. Sometimes this is possible, for instance when one is interested in growth curves of children's height. But in other circumstances, for instance a daily diary study of mood, one is hard pressed to say that the first occasion of measurement in the study actually has the same meaning for everyone. One person might start the study the day after winning the lottery and another the day after a car accident. Why would the experimenter wish to equate the meaning of time $t = 0$ for these two individuals? The differential equations model approach sidesteps this whole issue since it expresses a model that says if a person is in some particular state at a selected time t we predict that person will be changing at some instantaneous rate. By adding up these changes over an interval of time (integrating over an interval of time) we can predict an expected trajectory for that person. Thus, the differential equation model does not specify a particular trajectory. Rather, one may think of differential equations models as specifying a family of trajectories; each of which has a starting point, a set of states called *initial conditions* at some selected time $t = 0$.

Next, consider a bit more complicated model for change. Suppose that the slope of x is a function of the value of x so that

$$\dot{x}(t) = b_1 x(t) \tag{3}$$

This model for change says that the slope of x is proportional to the value of x. If $b_1 < 0$, then this model suggests that if we knew that x at some chosen time t was a positive value, then the slope would be a negative value. Thus as time progressed, x would approach 0. Similarly if $x(t)$ were negative, the slope would be positive and so as time progressed x would approach 0. Again, b_1 has units associated with it, for instance if x is a distance b_1 it might be expressed in meters per second. More to the point for behavioral variables, if x is in population standard deviation units and the experiment were a daily diary study, b_1 might be in units of standard deviations per day.

When the coefficient b_1 is negative in equation 3 the trajectory that is produced is a negative exponential. To see this, we can integrate equation 3 to find that

$$x(t) = b_0 e^{b_1 t} \tag{4}$$

Again, we find that we have an intercept term that enters into the picture, but this time as a scaling of the negative exponential function of time.

The expression of x as a nonlinear function of time in equation 4 may seem unfamiliar to many behavioral scientists. However, it is actually widely used in its discrete form without people realizing that their model implies a nonlinear function of time. Consider the following discrete time forward prediction model where

$$x(t + \Delta t) = c_1 x(t) \tag{5}$$

where Δt is the interval between occasions of measurement. This seemingly linear autoregressive model has a nonlinear component that is hidden in it.

To see this, consider the graphs in Figure 37.1(a), (b), and (c). These graphs appear identical to one another. Figure 37.1(a) was created by the differential equation in equation 3 choosing $x(0) = 4$ and $b_1 = -0.05$. Figure 37.1(b) was created by the differential equation in equation 4 choosing $b_0 = 4$ and $b_1 = -0.05$. Finally, the discrete measurements in Figure 37.1(b) were created by setting $x(0) = 4, \Delta t = 1$, and $c_1 = e^{(b_1 \Delta t)}$. Thus, the autoregressive parameter c_1 is a nonlinear function of the interval of time between successive measurements. The parameters of the differential and integral form are independent of the measurement interval.

In order to plot a single person's trajectory over time, we needed to choose an initial value

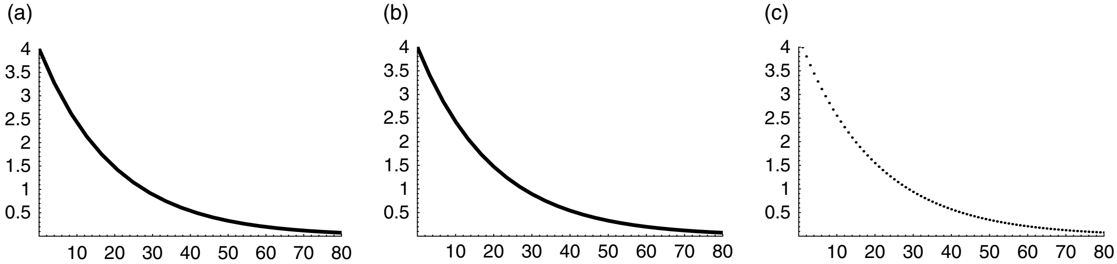


Figure 37.1 Negative exponential decay function as specified by (a) the differential equation in equation 3, (b) the integral form in equation 4, and (c) the discrete time form of autoregression in equation 5

for x at time 0. Thus, the differential form, the integral form, and the discrete form of these equations are equivalent when an initial value is known. If there are individual differences in initial values, the integral form and discrete form must be estimated as a random coefficients (i.e., mixed effects) model in order for the models to be equivalent. In this case the integral form or discrete form have the advantage of giving estimates of the interindividual differences in initial values, if that is important to the question at hand. If individual differences in initial values are unrelated to the research question, then the differential form simplifies the statistical model to be fit.

1 Second order equations

So far, we have considered only the simplest first order linear differential equations, equations that only involve the first derivative of a variable. Second order linear equations can exhibit many types of behavior, including exponential decline or increase, increase and then exponential decline, and oscillations as shown in Figure 37.2.

A second order linear differential equation in one variable may be written as

$$\ddot{x}(t) = b_1 x(t) + b_2 \dot{x}(t) \tag{6}$$

where $\ddot{x}(t)$ is the second derivative of x with respect to time. This equation expresses the expected simultaneous relationships between the time derivatives of x at one moment in time t. If both $b_1 < 0$ and $b_2 < 0$ this system is called a *damped linear oscillator*. If $b_1 < 0$ and $x(t) > 0$ the effect on the second derivative is negative, i.e., the slope is becoming more negative. Thus, the farther x is from 0 in a positive direction, the more the slope tends to become negative. The effect over time is that x is driven back towards the equilibrium value 0. Similarly, if $b_1 < 0$ and $x(t) < 0$, the effect on

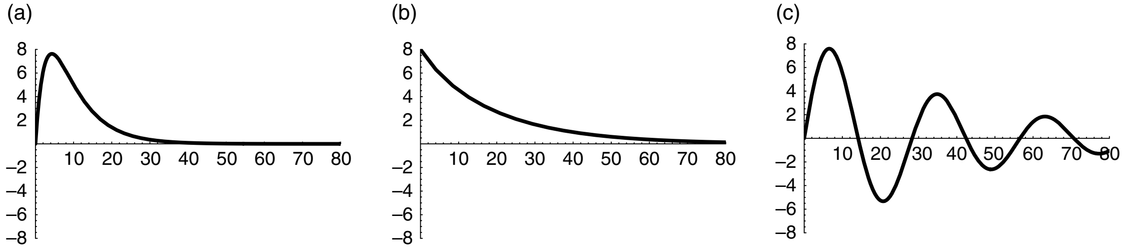


Figure 37.2 The same second order differential equation (equation 6) can produce (a) increase followed by decrease, (b) exponential decrease, and (c) oscillations depending on its parameters and initial values

the second derivative is positive and so the slope is becoming more positive and thus the result is that x is again driven back towards equilibrium. This oscillation would continue indefinitely except that if $b_2 < 0$ the greater the slope, the more the slope changes to be close to zero. When $b_2 < 0$ the system is said to *damp to equilibrium.*

There are many other forms of differential equations: higher order equations with third or fourth derivatives, equations that are nonlinear in their variables, equations that are nonlinear in their parameters, or equations that have time-dependent parameters (the interested reader is referred to Ellis, Johnson, Lodi and Schwalbe, 1992; Hubbard and West, 1991, 1995; Kaplan and Glass, 1995; Thompson and Stewart, 1986; Wylie, 1979). Furthermore, any of the preceding types of equations may be combined into coupled systems of equations: simultaneous equations whose time-dependent trajectories are interdependent on one another.

2 Some methods for estimating parameters

Once a model has been formulated in order to express the relationship between derivatives of a variable, the model can be fit to repeated observations data. There are a variety of techniques for estimating parameters of differential equations. These techniques fall into two main categories, integral forms and differential forms.

Differential equations estimation has a long history, tracing back to Hotelling (1927) who posed the problem of differential equations subject to error. The stochastic integral was introduced by Itô (1951) and this method has been furthered by several researchers (Bergstrom, 1966; Arminger, 1986) and applied in an exact discrete (Singer, 1993) and approximate discrete (Oud and Jansen, 2000) form to repeated observations data. These models have been shown to improve on cross-lag panel models by using the continuous time form of the first order differential equation equivalent (Oud, 2007). These methods use the integral form of the differential equation, so either an analytic integral or numerical approximation must be derived prior to fitting the model.

Another technique for fitting differential equation models in the integral form to time-series data involves using Kalman Filters (Kalman, 1960) or Extended Kalman Filters to fit a state–space linearization of the differential equation in question (for details see Harvey, 1989; Chatfield, 2004). These techniques show considerable promise for the flexible estimation of parameters of linear and nonlinear systems with time-varying parameters and have begun to be used in the behavioral sciences (Chow, 2006).

One discrete method that has been proposed casts the differential equation in terms of a *latent difference score* equation (Hamagami and McArdle, 2007; McArdle, 2000). This method uses latent variables to calculate difference scores in order to predict first order change. This method is a latent variable extension to first order autoregressive time-series methods that make linear predictions forward in time in a manner similar to equation 5. This method has the advantage of ease of specification and its latent difference score is easy to understand. However, its parameters are a nonlinear function of the differential equation parameters and the lag between measurements (recall that $c_1 = e^{b_1 \Delta t}$).

One differential form estimation two-step procedure was proposed by Boker and Graham (1998) and Boker and Nesselroade (2002) who used simplified local linear approximation to generate estimated derivatives and then fit the differential form of the model directly to the resulting derivatives. This method has been applied to coupled systems (Boker, 2001; Butner, Amazeen and Mulvey, 2005) and has been applied in a multilevel coupled context where individual differences in parameters are predicted by second level variables (Boker and

Laurenceau, 2005; Maxwell and Boker, 2007). While this method has an advantage of simplicity, its parameters are prone to bias if non-optimal time delays are not available (Boker and Nesselroade, 2002).

The remainder of the chapter focuses on another method for fitting the differential form of a differential equation to repeated observations: latent differential equations.

3 Latent differential equations

Latent differential equations (LDEs) use structural equation modeling to simultaneously estimate latent derivatives of a time series and fit a structural model to the covariances between those derivatives (Boker, Neale and Rausch, 2004). This method estimates measurement error, which does not affect the trajectory of the system over time, and dynamic error, which does affect the system's trajectory. The method is less prone to parameter bias over a range of time delays between occasions of measurement than is the local linear approximation method referred to previously.

LDE uses a form of Savitzky-Golay filtering (Savitzky and Golay, 1964) to construct a constrained loading matrix similar to that of a latent growth curve structural model. The data are time series that have been put into a time-delay embedded (i.e. state space) matrix with five time delay lags on every row of the matrix. The method for constructing this matrix is described in detail in the next section. The covariance structure between the latent variables of this model can be used to specify a differential equation model as in, for instance, equations 3 or 6.

3.1 Time-delay embedding

Most methods that estimate differential equations models use some form of state space embedding. Time-delay embedding constructs a form of state space in which the data columns are time lagged. The simplest form of a time-delay embedded matrix is a pre-post design where the first column of the matrix is the observation before an intervention and the second column is the same variable measured some time Δt after the first observation, and presumably after the intervention. Each row in this matrix is a different individual, and the assumption is made that the individual differences in the change between column 1 and column 2 provide a measure of the process of interest.

Suppose there are five equal interval repeated observations per person. The data may again be arranged so that each row contains observations from one individual and the time delay between columns is again Δt. Each row of the resulting matrix thus represents a total interval of time equal to the number of columns minus one times the delay between columns, in our case $(5-1)\Delta t$. Again, the assumption is made that the process evolving during the time that elapses in the interval between the occasions of measurement in the first column and the last column in the row will manifest itself in the relationship between the data in the 5 columns. It is assumed that to the degree that individuals in the sample are representative of the population, statistics calculated using the relationships between columns will be representative of the process of interest. Under assumptions of ergodicity, we can use a time-delay embedding to estimate the parameters of differential equation models (for details see Noakes, 1991; Sauer, Yorke and Casdagli, 1991; Takens, 1985; Whitney, 1936).

Now, suppose that we have 100 observations per person. If the relationships between the 5 columns in the previous matrix were sufficient to capture the time evolution of the process, we do not need more columns in our data matrix. But since we have 100 observations per person we wish to use these data effectively. If the data for person 1 on occasions 1 through 5 is $\{x(1,1), x(1,2), x(1,3), x(1,4), x(1,5)\}$, then we could consider that to be one observation of the time evolution of the process of interest. Another observation from the same person would be $\{x(1,2), x(1,3), x(1,4), x(1,5), x(1,6)\}$. In this way we could march through the first person's data constructing $100-4=96$ observations of the time evolution of the process. Thus we can add 96 rows to our time-delay matrix for each 100 observations per person.

But what if the total delay in each row $(4\Delta t)$ is not long enough to capture the change in which we are interested? It might be that the process evolves relatively slowly and a longer delay is necessary to observe the change. The Takens (1985) embedding theorem says that the time-delay embedding technique is sufficient if enough columns are chosen and the time between columns is not poorly chosen. In practice, there will be a time delay between the first and last column that will work best, maximizing the ratio of the reliable change over the total change. When we have many measurements, such as the 100 measurements postulated above, we can vary the interval between columns by selecting an occasion indexing parameter τ that gives the number of occasions to lag between each column. In the example in the previous paragraph we used $\tau = 1$. In general the system is unknown, and so τ is frequently selected over a range of values and models fit to embedded data from each selected value of τ so as to test the stability of parameter estimates.

Consider a univariate time series X where individuals $i = 1 \ldots N$ are observed on occasions $j = 1 \ldots P$ separated by a fixed time interval Δt. We will create a $d = 5$ dimensional embedding (5 lagged data columns) such that the time delay between embedded columns is $\tau\Delta t$ where $\tau = 2$. The embedding delay will thus be twice the interval between occasions of measurement. If the original time series X is ordered by occasion j within individual i then the series of $x_{(i,j)}$ can be written as a vector of scores

$$X = \{x_{(1,1)}, \ldots\ x_{(1,P)},\ x_{(2,1)}, \ldots\ x_{(2,P)}, \ldots\ x_{(N,1)}, \ldots\ x_{(N,P)}\} \qquad (7)$$

The five dimensional $(d = 5)$ embedding $X^{(5)}$ where $\tau = 2$ can then be written as a matrix with five columns such that

$$X^{(5)} = \begin{bmatrix}
x_{(1,1)} & x_{(1,3)} & x_{(1,5)} & x_{(1,7)} & x_{(1,9)} \\
x_{(1,2)} & x_{(1,4)} & x_{(1,6)} & x_{(1,8)} & x_{(1,10)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x_{(2,P-8)} & x_{(2,P-6)} & x_{(2,P-4)} & x_{(2,P-2)} & x_{(2,P)} \\
x_{(2,1)} & x_{(2,3)} & x_{(2,5)} & x_{(2,7)} & x_{(2,9)} \\
x_{(2,2)} & x_{(2,4)} & x_{(2,6)} & x_{(2,8)} & x_{(2,10)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x_{(2,P-8)} & x_{(2,P-6)} & x_{(2,P-4)} & x_{(2,P-2)} & x_{(2,P)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x_{(N,1)} & x_{(N,3)} & x_{(N,5)} & x_{(N,7)} & x_{(N,9)} \\
x_{(N,2)} & x_{(N,4)} & x_{(N,6)} & x_{(N,8)} & x_{(N,10)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x_{(N,P-8)} & x_{(N,P-6)} & x_{(N,P-4)} & x_{(N,P-2)} & x_{(N,P)}
\end{bmatrix} \qquad (8)$$

Each row of $X^{(5)}$ is a short within-person sequence of observations where an interval of $\tau\Delta t$ separates each column. Each individual's data are used fully as long as there are at least $P \geq 10$ occasions of measurement for each individual. Finally, note that the data from individual i never appears on the same row with individual $i+1$; thus data from one person's process does not overlap with data from another individual.

In order to construct a time-delay embedding we must determine how many columns are in the embedding. For use with an LDE model, one will frequently use between $d = 4$ and $d = 6$ columns. Four columns is the minimum number of columns that is still identified for a second order differential equation model. Five columns has, in simulations, proved to be somewhat more stable in the case of a second order linear differential equation, and allows the estimation of models other than the simplest second order model. More than $d = 6$ columns can be used, but in any model that includes the possibility of oscillation one must take care that the elapsed time between the first column and the last column in the embedding does not exceed one-half the elapsed time between peaks of the oscillation. Thus, $d\tau\Delta t$ must be less than one-half the period of the oscillation if the LDE method is to be applied to oscillating data.

3.2 First order LDE

In the next sections we will fit linear first and second order differential equations as were specified earlier. But now we will add a residual term so that for the first order linear differential equation we have

$$\dot{x}_t = b_1 x_t + e_t \tag{9}$$

where $\dot{x}_t$ is the first derivative of x with respect to time at time t, x_t is the displacement of the variable x from its equilibrium value at time t, and e_t is an approximately normally distributed independent residual. We can use a slope and intercept latent growth model with fixed loadings to estimate the parameter of interest, b_1, from a five-dimensional time-delay embedded data matrix as shown in the path diagram in Figure 37.3.

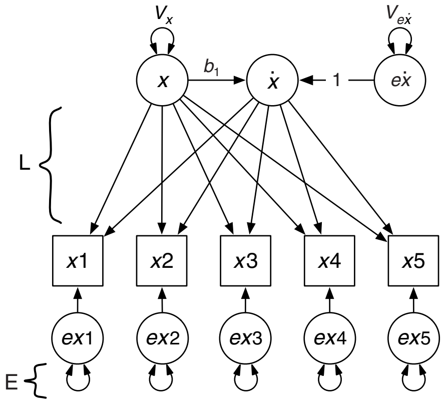


Figure 37.3 Path diagram of a first order linear differential equation specified as an LDE model with five indicators (i.e., a five-dimensional time-delay embedding of the time-series data

The loading matrix **L** will, in this case, be of order 5 × 2 and is specified so that the first column is a column of 1's and the second column is a linear basis function scaled by the interval between the columns, $\tau\Delta t$, and centered on the middle column of the time-delay embedded matrix.

$$\mathbf{L} = \begin{bmatrix} 1 & -2\tau\Delta t \\ 1 & -\tau\Delta t \\ 1 & 0 \\ 1 & \tau\Delta t \\ 1 & 2\tau\Delta t \end{bmatrix} \tag{10}$$

In this way, the latent variables x and $\dot{x}$ in Figure 37.3 will be the intercept and slope centered around the observation $x3$, the third column of the five-dimensional time-delay embedded data matrix. The structural part of the model can be specified using RAM matrices (McArdle and McDonald, 1984) as

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ b_1 & 0 \end{bmatrix} \tag{11}$$

$$\mathbf{S} = \begin{bmatrix} V_x & 0 \\ 0 & V_{e\dot{x}} \end{bmatrix} \tag{12}$$

If we now define a 5 × 5 diagonal matrix **E** to contain the five residual variances for the five indicators (V_{ex1}, V_{ex2}, V_{ex3}, V_{ex4}, V_{ex5}), we can calculate the expected covariance matrix **R** for this model as

$$\mathbf{R} = \mathbf{L}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{S}(\mathbf{I}-\mathbf{A})^{-1'}\mathbf{L}' + \mathbf{E} \tag{13}$$

The estimate for b_1 corresponds to the parameter in equation 9 while the estimate for $V_{e\dot{x}}$ corresponds to the variance of the residual e_t. Note that the residual variances in **E** are the portion of the variance of the indicators that does not conform to the latent model where only an intercept and slope are used to account for the time dependence in each row of the time-delay embedded data matrix. The model may be misspecified if higher order derivatives are required to account for this time dependence. The variance $V_{e\dot{x}}$ estimates the portion of the variance of the slope that cannot be accounted for by the displacement from equilibrium, i.e., the intercept. This residual provides an estimate of the slope variance that does not conform to a linear first order differential equation. To the extent that $V_{e\dot{x}}$ is large, it may be that there are exogenous influences with effects that propagate linearly over time. Or it may be that the linear relationship between x and $\dot{x}$ is insufficient to capture the relationship between displacement from equilibrium and its first derivative.

3.3 Second order LDE

In this section we fit a linear second order differential equation specified as

$$\ddot{x}_t = \eta x_t + \zeta \dot{x}_t + e_t \tag{14}$$

where $\dot{x}_t$ and $\ddot{x}_t$ are the first and second derivatives respectively of x with respect to time at a particular time t, x_t is the displacement of the variable x from its equilibrium value at time t, and e_t is an approximately normally distributed independent residual. The coefficients η and ζ are frequency and damping coefficients respectively of the damped linear oscillator formed when $\eta < 0$. This model is one of the simplest methods for accounting for selfregulating systems that have a stable equilibrium or *set point*. The farther the system is from equilibrium (when x_t is large), the more the system curves ($\ddot{x}_t$ becomes large and of opposite sign to x_t) back towards equilibrium. The larger a negative number is η, the greater this effect. One might consider this as the system attempting to avoid being far from equilibrium: the farther it is from equilibrium, the more it curves and goes back. Similarly, one can think of $\zeta < 0$ as a coefficient controlling the avoidance of rapid change: the faster the system is changing, the more it decelerates.

Suppose we have created a five-dimensional time-delay embedding matrix $X^{(5)}$ as described above. Each of the column vectors of this matrix is a manifest variable indicator of the second order differential equation shown in Figure 37.4. The latent variables in this model are estimates of the displacement, first derivative, and second derivative of each row of $X^{(5)}$. A three-factor confirmatory model with five indicators is not identified when the factor loadings are allowed to be free. In the case of the LDE model, the loading matrix **L** is constrained so as to estimate latent derivatives.

The LDE loading matrix **L** is a number of indicators by number of latent variables matrix whose values are constrained as follows. The first column of the matrix is fixed to be equal to one. The second column is fixed to be a unit basis function for a slope scaled by the interval between columns and with intercept at the middle indicator. The third column is the indefinite integral of the second column; i.e., the third column is the second column squared and divided by two. Higher order derivatives can be calculated in the same way. Each successive column in **L** will be the indefinite integral of the previous column. Thus in the case of a second order LDE model of a five-dimensional embedding, we can write **L** as

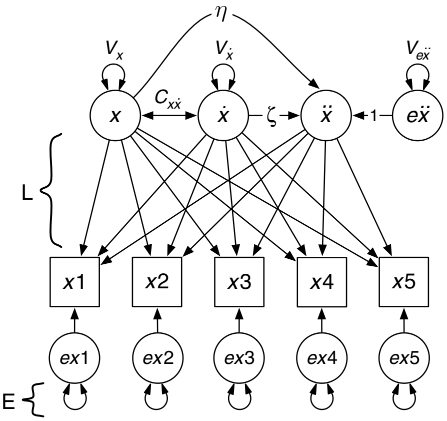


Figure 37.4 Path diagram of a second order latent differential equation model

$$\mathbf{L} = \begin{bmatrix} 1 & -2\tau\Delta t & (-2\tau\Delta t)^2/2 \\ 1 & -\tau\Delta t & (-\tau\Delta t)^2/2 \\ 1 & 0 & 0 \\ 1 & \tau\Delta t & (\tau\Delta t)^2/2 \\ 1 & 2\tau\Delta t & (2\tau\Delta t)^2/2 \end{bmatrix} \quad (15)$$

where Δt is the elapsed time between successive occasions of measurement and τ is the number of occasions of measurement separating each column of the time-delay embedded matrix.

The covariances between the displacement x, the latent first derivative $\dot{x}$ and the latent second deriviative $\ddot{x}$ are used to estimate the parameters of a second order differential equations model in equation 14. The regression coefficients η and ζ appear in a 3×3 matrix $\mathbf{A}$ and the free variances and covariances of the latent variables appear in a matrix $\mathbf{S}$.

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \eta & \zeta & 0 \end{bmatrix} \quad (16)$$

$$\mathbf{S} = \begin{bmatrix} V_x & C_{x\dot{x}} & 0 \\ C_{x\dot{x}} & V_{\dot{x}} & 0 \\ 0 & 0 & V_{e\ddot{x}} \end{bmatrix} \quad (17)$$

Again we define a 5×5 diagonal matrix $\mathbf{E}$ as in the previous section and estimate the expected covariance matrix $\mathbf{R}$ for this model as

$$\mathbf{R} = \mathbf{L}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{S}(\mathbf{I}-\mathbf{A})^{-1'}\mathbf{L}' + \mathbf{E} \quad (18)$$

The five columns of $X^{(5)}$ gives us 15 degrees of freedom in the data covariances. There are five degrees of freedom used by $\mathbf{E}$, two degrees of freedom used by $\mathbf{A}$ and five degrees of freedom used by $\mathbf{S}$, leaving four degrees of freedom with which to test the fit of the model.

4 Multivariate second order LDE

The LDE model specified in the previous section was only indicated by a single variable, although its time structure was converted into a multivariate form by using time-delay embedding. A construct indicated by multiple manifest variables may also be time-delay embedded so as to create a data matrix that has multivariate indicators each of which is multivariate across time. This form of time-delay embedding produces a better estimate of the dynamics of a latent construct since the differential equation coefficients are estimated using only the common variance that is time structured.

Constructing a multivariate time-delay embedding is a straightforward extension of the time-delay embedding procedure described above. Suppose we have three time series X, Y, and Z where individuals $i = 1 \ldots N$ are observed on occasions $j = 1 \ldots P$ separated by a fixed time interval s. Again we will create a $d = 5$ dimensional embedding such that the time delay between embedded columns is τs where $\tau = 2$, thus the embedding delay is twice the interval between occasions of measurement. If the original time series X, Y, and Z are ordered by occasion j within individual i, then these series can be written as vectors of scores

$$\begin{aligned} X =& \{x_{(1,1)}, \ldots x_{(1,P)}, x_{(2,1)}, \ldots x_{(2,P)}, \ldots, \\ & x_{(N,1)}, \ldots x_{(N,P)}\} \\ Y =& \{y_{(1,1)}, \ldots y_{(1,P)}, y_{(2,1)}, \ldots y_{(2,P)}, \ldots, \\ & y_{(N,1)}, \ldots y_{(N,P)}\} \\ Z =& \{z_{(1,1)}, \ldots z_{(1,P)}, z_{(2,1)}, \ldots z_{(2,P)}, \ldots, \\ & z_{(N,1)}, \ldots z_{(N,P)}\} \end{aligned} \quad (19)$$

We first construct three embedding matrices $X^{(5)}, Y^{(5)}$, and $Z^{(5)}$ as shown previously in equation 8. We then augment these three matrices together $(X^{(5)}|Y^{(5)}|Z^{(5)})$ so that the rows align. For example, when $\tau = 2$, the augmented time-delay embedded matrix would take the form

$$
W^{(5)} = \begin{bmatrix}
x_{(1,1)} & \cdots & x_{(1,9)} & y_{(1,1)} & \cdots & y_{(1,9)} & z_{(1,1)} & \cdots & z_{(1,9)} \\
x_{(1,2)} & \cdots & x_{(1,10)} & y_{(1,2)} & \cdots & y_{(1,10)} & z_{(1,2)} & \cdots & z_{(1,10)} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_{(2,P-8)} & \cdots & x_{(2,P)} & y_{(2,P-8)} & \cdots & y_{(2,P)} & z_{(2,P-8)} & \cdots & z_{(2,P)} \\
x_{(2,1)} & \cdots & x_{(2,9)} & y_{(2,1)} & \cdots & y_{(2,9)} & z_{(2,1)} & \cdots & z_{(2,9)} \\
x_{(2,2)} & \cdots & x_{(2,10)} & y_{(2,2)} & \cdots & y_{(2,10)} & z_{(2,2)} & \cdots & z_{(2,10)} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_{(2,P-8)} & \cdots & x_{(2,P)} & y_{(2,P-8)} & \cdots & y_{(2,P)} & z_{(2,P-8)} & \cdots & z_{(2,P)} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_{(N,1)} & \cdots & x_{(N,9)} & y_{(N,1)} & \cdots & y_{(N,9)} & z_{(N,1)} & \cdots & z_{(N,9)} \\
x_{(N,2)} & \cdots & x_{(N,10)} & y_{(N,2)} & \cdots & y_{(N,10)} & z_{(N,2)} & \cdots & z_{(N,10)} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_{(N,P-8)} & \cdots & x_{(N,P)} & y_{(N,P-8)} & \cdots & y_{(N,P)} & z_{(N,P-8)} & \cdots & z_{(N,P)}
\end{bmatrix} \tag{20}
$$

Thus, the first five columns of $W^{(5)}$ are the time-delay embedded values for the variable x, the next five columns for y, and the last five columns for z. This data matrix can now be fit by the multivariate LDE models, such as the second order linear model shown in Figure 37.5.

Since the multivariate LDE model is estimating both the within-time factor structure of the latent construct F as well as the time-delayed structure as derivatives, loading matrix will no longer be entirely constrained to fixed values. The first five rows of $\mathbf{L}$ are the same as in the univariate case: all values are fixed since both Δt and τ are known in advance. The sixth through tenth rows of $\mathbf{L}$ are simply a copy of the first five rows scaled by the free coefficient a. Similarly, the eleventh through fifteenth rows are scaled by the free coefficient b resulting in

$$
\mathbf{L} = \begin{bmatrix}
1 & -2\tau\Delta t & (-2\tau\Delta t)^2/2 \\
1 & -\tau\Delta t & (-\tau\Delta t)^2/2 \\
1 & 0 & 0 \\
1 & \tau\Delta t & (\tau\Delta t)^2/2 \\
1 & 2\tau\Delta t & (2\tau\Delta t)^2/2 \\
a & -2a\tau\Delta t & a(-2\tau\Delta t)^2/2 \\
a & -a\tau\Delta t & a(-\tau\Delta t)^2/2 \\
a & 0 & 0 \\
a & a\tau\Delta t & a(\tau\Delta t)^2/2 \\
a & 2a\tau\Delta t & a(2\tau\Delta t)^2/2 \\
b & -2b\tau\Delta t & b(-2\tau\Delta t)^2/2 \\
b & -b\tau\Delta t & b(-\tau\Delta t)^2/2 \\
b & 0 & 0 \\
b & b\tau\Delta t & b(\tau\Delta t)^2/2 \\
b & 2b\tau\Delta t & b(2\tau\Delta t)^2/2
\end{bmatrix} \tag{21}
$$

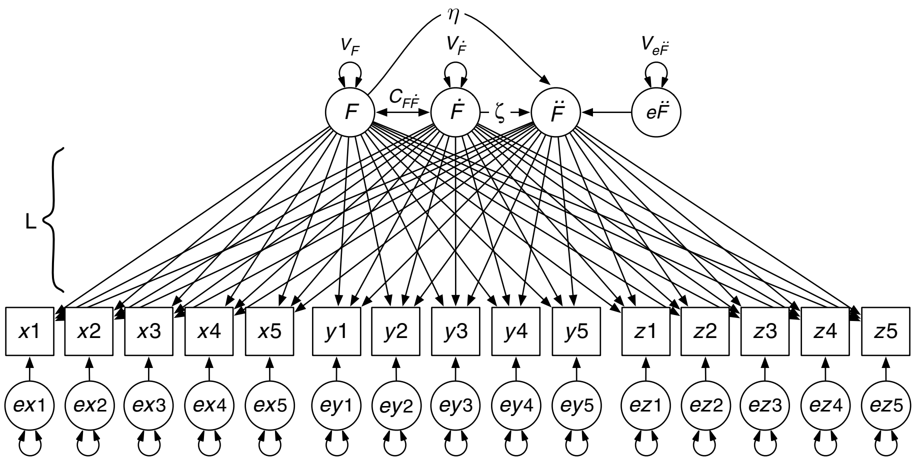


Figure 37.5 Path diagram of a multivariate second order latent differential equation model with three indicators and a five-dimensional time-delay embedding

When **L** is constructed in this manner, a and b will be factor loadings for the construct such that the factor structure is invariant over time and over the derivatives of F. We call this *differential factor invariance*. Differential invariance is expected to hold when the differential equations model is linear in its coefficients and variables. This hypothesis can be tested by releasing the constraint that the factor loadings a and b be equal across the columns of **L**.

The structural part of this multivariate second order linear LDE is exactly the same as the univariate case from equations 16 and 17 shown above, although we have now labeled our construct as F.

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \eta & \zeta & 0 \end{bmatrix} \tag{22}$$

$$\mathbf{S} = \begin{bmatrix} V_F & C_{F\dot{F}} & 0 \\ C_{F\dot{F}} & V_{\dot{F}} & 0 \\ 0 & 0 & V_{\ddot{F}} \end{bmatrix} \tag{23}$$

Again, the predicted covariance of the multivariate model can be calculated as shown in equation 18.

5 LDE model extensions

The latent structure of LDE models is flexible with respect to how the differential equation or equations are specified. For instance, a fourth order model can be created by adding two extra columns to the time-delay embedded data matrix, and setting up the **L** matrix to have 7 rows and 5 columns (Boker, 2007). Or coupled differential equations can be created by augmenting two time-delay embedded matrices together and setting up a structural model where two latent variables are intrinsically regulated as well as bidirectionally coupled:

$$\begin{aligned}\ddot{x}(t) = \eta_x x(t) + \zeta_x \dot{x}(t) + \gamma_y(\eta_y y(t) \\ + \zeta_y \dot{y}(t)) + e_{\ddot{x}}(t)\end{aligned} \tag{24}$$

$$\begin{aligned}\ddot{y}(t) = \eta_y y(t) + \zeta_y \dot{y}(t) + \gamma_x(\eta_x x(t) \\ + \zeta_x \dot{x}(t)) + e_{\ddot{y}}(t)\end{aligned} \tag{25}$$

The second derivative of y and x are regulated both by their own intrinsic process, but also as a linear proportion of the other variable's regulation. Systems like these have been used to model dyadic relationships in married couples (Boker and Laurenceau, in press).

6 Limitations and recommendations

There are limitations to the methods described in this chapter. The first limitation is that there must be intensive longitudinal measurement in order to estimate parameters of all but the simplest of differential equations. As the differential equation models become more complex, the need for intensive measurement per individual increases. Nonlinear models in particular can require hundreds or even thousands of measurements per individual. While this is within the realm of possibility for physiological data, thousands of questionnaires per individual is beyond the scope of behavioral research.

LDE and time-delay embedding can be used with as few as five observations per person, but unless there are large interindividual differences in regulation or the model is a simple first order model, greater power is gained from five additional measurements of that individual than is gained by adding a new individual with five measurements. The reason is that the new individual only adds one observation of the relationship between the columns, but adding five observations to an existing individual adds five rows to the time-delay embedding matrix. For the same number of measurements, intensive longitudinal measurement almost always will be more powerful than short bursts when testing differential equations models.

Currently, standard errors for LDE models estimated parametrically from SEM packages are incorrect. Nonindependence of rows in time-delay embedded data matrices violate the assumptions on which these standard errors are calculated. Structured bootstrap methods are currently under development in order to address this problem. Until these methods have been verified, nested model comparison using fit statistics is recommended.

7 Conclusions

The methods reviewed and presented in this chapter are relatively new in behavioral research. These methods can be usefully applied, but require more explanation than tried and true longitudinal methods. This places a burden on researchers reporting results from these methods to describe the implications of their models in meaningful theoretic terms. However, differential equations models can lead to insights in behavioral regulation that may not be apparent when using other statistical techniques.

Differential equations provide a way to frame theories of regulation and time dependency that are flexible and result in parameters that have useful theoretic interpretations. It can be helpful to think through one's theory in terms of, "What causes change?" and "How does this system regulate itself as opposed to being regulated by some extrinsic variable?" Framing one's theory in these terms has led naturally to specifying and fitting differential equations models for emotional regulation, ovarian hormones coupled to disorder eating, perception-action systems in posture, cognitive aging, interpersonal communication, and social interaction. It is expected that differential equations models will find their way into wide usage in the behavioral sciences in the coming decades.

Author note

Funding for this work was provided in part by NSF Grant BCS–0527485. Correspondence may be addressed to Steven M. Boker, Department of Psychology, The University of Virginia, PO Box 400400, Charlottesville, VA 22903, USA; email sent to smb3u@virginia.edu; or browsers pointed to http://www.people.virginia.edu/~smb3u.

References

Arminger, G. (1986). Linear stochastic differential equation models for panel data with unobserved variables. In N. Tuma (ed.), *Sociological Methodology*, pp. 187–212. San Francisco: Jossey Bass.

Bergstrom, A. R. (1966). Nonrecursive models as discrete approximations to systems of stochastic differential equations. *Econometrica*, 34: 173–182.

Boker, S. M. (2001). Differential structural modeling of intraindividual variability. In L. Collins and A. Sayer (eds), *New Methods for the Analysis of Change*, pp. 3–28. Washington, DC: APA.

Boker, S. M. (2007). Specifying latent differential equations models. In S. M. Boker and M. J. Wenger (eds), *Data Analytic Techniques for Dynamical Systems*, pp. 131–158. Mahwah, NJ: Lawrence Erlbaum.

Boker, S. M. and Graham, J. (1998). A dynamical systems analysis of adolescent substance abuse. *Multivariate Behavioral Research*, 33(4): 479–507.

Boker, S. M. and Laurenceau, J. P. (2005). Dynamical systems modeling: An application to the regulation of intimacy and disclosure in marriage. In T. A. Walls and J. L. Schafer (eds), *Models for Intensive Longitudinal Data*, pp. 195–218. Oxford: Oxford University Press.

Boker, S. M. and Laurenceau, J. P. (in press). Coupled dynamics and mutually adaptive context. In T. D. Little, J. A. Bovaird and N. A. Card (eds), *Modeling Ecological and Contextual Effects in Longitudinal Studies of Human Development*. Mahwah, NJ: Lawrence Erlbaum.

Boker, S. M., Neale, M. C. and Rausch, J. (2004). Latent differential equation modeling with multivariate multioccasion indicators. In K. van Montfort, H. Oud and A. Satorra (eds), *Recent Developments on Structural Equation Models: Theory and Applications*, pp. 151–174. Dordrecht, Netherlands: Kluwer Academic.

Boker, S. M. and Nesselroade, J. R. (2002). A method for modeling the intrinsic dynamics of intraindividual variability: Recovering the parameters of simulated oscillators in multiwave panel data. *Multivariate Behavioral Research*, 37(1): 127–160.

Butner, J., Amazeen, P. G. and Mulvey, G. M. (2005). Multilevel modeling to two cyclical processes: Extending differential structural equation modeling to nonlinear coupled systems. *Psychological Methods*, 10(2): 159–177.

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*. New York: CRC Press.

Chow, S. M. (2006). Factor score and parameter estimation in nonlinear dynamical systems models. In K. van Montfort and A. Satorra (eds), *Longitudinal Models in the Behavioral and Related Sciences (Vol. 2)*. Mahwah, NJ: Lawrence Erlbaum.

Ellis, W., Johnson, E., Lodi, E. and Schwalbe, D. (1992). *Maple V Flight Manual: Tutorials for Calculus, Linear Algebra and Differential Equations*. Pacific Grove, CA: Brooks/Cole.

Hamagami, F. and McArdle, J. J. (2007). Dynamic extensions of latent difference score models. In S. M. Boker and M. J. Wenger (eds), *Data Analytic Techniques for Dynamical Systems*, pp. 47–86. Mahwah, NJ: Lawrence Erlbaum.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Princeton, NJ: Princeton University Press.

Hotelling, H. (1927). Differential equations subject to error, and population estimates. *Journal of the American Statistical Association*, 22(159): 283–314.

Hubbard, J. H. and West, B. H. (1991). *Differential Equations: A Dynamical Systems Approach*. New York: Springer-Verlag.

Hubbard, J. H. and West, B. H. (1995). *Differential Equations: A Dynamical Systems Approach, Higher Dimensional Systems*. New York: Springer-Verlag.

Itô, K. (1951). On stochastic differential equations. *American Mathematical Society Memoirs (No. 4)*.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering*, 82: 35–45.

Kaplan, D. and Glass, L. (1995). *Understanding Nonlinear Dynamics*. New York: Springer-Verlag.

Maxwell, S. E. and Boker, S. M. (2007). Multilevel models of dynamical systems. In S. M. Boker and M. J. Wenger (eds), *Data Analytic Techniques for Dynamical Systems*, pp. 161–188. Mahwah, NJ: Lawrence Erlbaum.

McArdle, J. J. (2000). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. du Toit and D. Sorbom (eds), *Structural Equation Modeling: Present and Future*, pp. 342–380. Lincolnwood, IL: Scientific Software International.

McArdle, J. J. and McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, 87: 234–251.

Noakes, L. (1991). The Takens embedding theorem. *International Journal of Bifurcation and Chaos*, 4(1): 867–872.

Oud, J. H. L. (2007). Continuous time modeling of reciprocal relationships in the cross-lagged panel design. In S. M. Boker and M. J. Wenger (eds), *Data Analytic Techniques for Dynamical Systems*, pp. 87–130. Mahwah, NJ: Lawrence Erlbaum.

Oud, J. H. L. and Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrica*, 65(2): 199–215.

Sauer, T., Yorke, J. and Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*, 65(3,4): 95–116.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares. *Analytical Chemistry*, 53: 1627–1639.

Singer, H. (1993). Continuous-time dynamical systems with sampled data, errors of measurement and unobserved components. *Journal of Time Series Analysis*, 14: 527–545.

Takens, F. (1985). Detecting strange attractors in turbulence. In A. Dold and B. Eckman (eds), *Lecture Notes in Mathematics 1125: Dynamical Systems and Bifurcations*, pp. 99–106. Berlin: Springer-Verlag.

Thompson, J. M. T. and Stewart, H. B. (1986). *Nonlinear Dynamics and Chaos*. New York: Wiley.

Whitney, H. (1936). Differentiable manifolds. *Annals of Mathematics*, 37: 645–680.

Wylie, C. R. (1979). *Differential Equations*. New York: McGraw-Hill.

Chapter 38

Nonlinear dynamics, chaos, and catastrophe theory

Courtney Brown

Nonlinear dynamics refers to a broad range of behavior that can occur with many varieties of mathematical models that are structured with respect to time. Interest in nonlinear dynamics has increased in recent years in the social sciences, in large part fueled by frustrations with the limitations associated with linear regression models. Nonlinear dynamics can occur with dynamic models that are algebraically linear or nonlinear. For example, both of the models $dy/dt = ay$ and $dy/dt = ay^2$ (where the parameter "a" is a constant) express nonlinear longitudinal behavior, yet the first is algebraically linear whereas the second is algebraically nonlinear. Minimally, nonlinear dynamics require that change occurs over time with respect to one or more variables such that this change cannot be represented as a straight line on a graph that places time on the horizontal axis. More specifically, nonlinear variation is on-going change that is not a constant increment with respect to time. Chaos and catastrophe theory are two subcategories within the area of nonlinear dynamics that address particular types of highly nonlinear behaviors peculiar to certain classes of functionally nonlinear dynamic models.

1 Nonlinear dynamics

The term "nonlinear dynamics" is often used within the context of complex situations in which phenomena with longitudinal change are modeled using nonlinear algebraic formulations (often involving systems of interdependent equations) that either implicitly or explicitly reference time. However, readers should note that linear dynamic models that exhibit nonlinear overtime behaviors have historically played an important role in the development of our contemporary understanding of nonlinear dynamics, including the nonlinear dynamics associated with algebraically nonlinear models. The two-nation linear arms race model of Lewis Fry Richardson is a prominent example of this (Richardson, 1960). A thorough mathematical introduction to nonlinear dynamics as it appears in both linear and nonlinear mathematical specifications can be found in Hirsch, Smale and Devaney (2003; also see Hirsch and Smale, 1974). Treatments and examples of this same subject matter from a social science perspective can be found in Brown (2007a, 1995a, 1995b, 1991). Also of interest, a seminal work edited by Diana Richards (2000) presents a collection of papers representing a large variety of nonlinear applications in the social sciences.

In the physical and natural sciences, nonlinear dynamics are normally encountered using continuous-time differential equation model specifications, which is a consequence of continuous-time processes of change that are naturally encountered in these fields. Exceptions are not rare, however, especially in the biological sciences, in which generational or seasonal changes are the focus of study. In instances, difference equations are sometimes employed. In the social sciences, nonlinear dynamics are often modeled using difference equations, almost always due to the manner in which social scientific data are collected (e.g., periodic elections, decade-spaced census reports, periodically-spaced survey data, etc.). Exceptions do occur here as well, and the technical difficulties involved in using continuous-time models with discretely measured data sets as are commonly encountered in the social sciences now have clear solutions (e.g., see Brown, 1995b). The choice of using either a continuous- or a discrete-time approach to modeling social phenomena has substantive consequences that can be important in certain settings, and a discussion of these substantive consequences can be found in Brown (1995b, pp. 13–30).

Nonlinear dynamics are normally described in terms of the following processes of change: (1) regular, (2) periodic, (3) chaotic, and (4) catastrophe. A regular process begins with a bifurcation, or a point in which a new process of change begins that differs structurally from a previous process of change. This bifurcation is normally followed by growth that initially experiences positive feedback that later changes to negative feedback as the growth process slows. The point at which positive feedback switches to negative feedback is called an inflection point, and in simple models this is where the first derivative switches in sign from positive to negative. The growth eventually tapers off entirely as the model asymptotically approaches an equilibrium, which is a constant steady state. At equilibrium, there is zero net change. Because of other processes of change that are temporarily external to this regular process, the regular process eventually becomes “ripe” for experiencing another bifurcation, which can lead to the initiation of a new process of change that can be any of the four listed above (including a new regular process).

A classic example of a regular process involving nonlinear dynamics is the logistic equation $dy/dt = ay(k-y)$, where growth in the variable y continues smoothly as its values asymptotically approach the equilibrium value k. This model is represented here as Figure 38.1, and the equilibrium value in this figure is 1.6. This model is functionally nonlinear since it contains a power term (y^2). The nonlinear dynamics aspect of the model is evidenced by the S-shape of the curve over time.

A periodic process is one in which the values of the relevant variables recur at specified intervals. Periodic processes normally have periodic limit cycles, as compared with fixed-point equilibria that are typically associated with regular processes. Given the fact that humans live lives

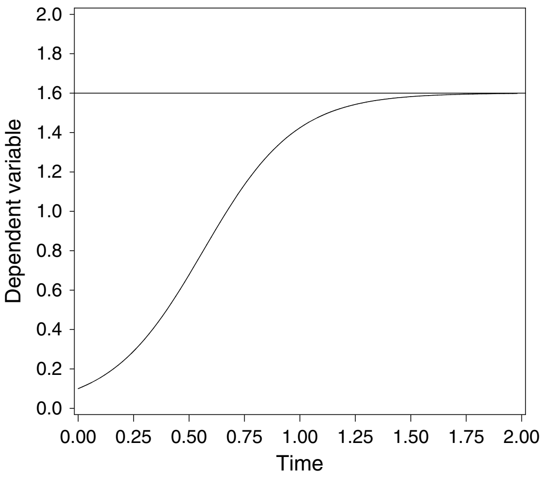


Figure 38.1 The logistic curve

according to countless cycles, it is surprising that periodic influences are not more commonly studied in the social sciences (relative to other types of analyses using, say, linear regression to find correlations between variables). For example, humans go to sleep and wake up periodically. We go to school in seasons. We eat with regular periodicity, breakfast, lunch, and dinner. We participate in elections in regular intervals. Our censuses are conducted in regular intervals. We pay our taxes yearly, and on and on.

A useful example of a periodic process is voting for the US Congress. Every two years there is a congressional election. But every four years there is also a presidential election. Interest in the presidential election increases turnout for the congressional contest every four years. This results in a surge in voter mobilization for Congress every four years in the so-called "on-year" elections. Subsequent elections two years later (the "off-year" elections) experience a decline in voter turnout due to the absence of a simultaneous presidential contest. If one were to model this, one might use a linear functional form originally suggested to me by John Sprague,

$$M_{t+1} = aM_t + b \tag{1}$$

where M_t is proportion of the eligible electorate that votes at time t, and a and b are parameters of the model. This is a first-order linear difference equation with constant coefficients.

For equation (1) to demonstrate the type of periodicity that is required given the structure of the US electoral calendar, parameter a must equal -1. That is the only value for that parameter that will produce finite oscillations of the type required for this electoral setting. Thus, we know the value of parameter a in advance without having to estimate it, and the only parameter that can be estimated for this model is parameter b. If one estimates this model using US congressional mobilization data from 1950 to 1970, then parameter $b = 0.9931$, and $R^2 = 0.68$.

The above model is best interpreted graphically. Figure 38.2 is constructed by first using the above values for the parameters a and b to calculate a predicted time series for M_t with respect to equation (1). Then these values of M_t are plotted on top of the actual data. The historical data are represented as dots, and the model is represented as the saw-toothed line.

Note in Figure 38.2 that the finite oscillations of the model closely correspond with the oscillatory characteristic of the data. Note also that this saw-toothed line clearly portrays the underlying periodic nature of these data. Remember that this model is linear in its functional form, yet it is still capable of expressing nonlinear dynamics of a periodic nature. Note also that the data for this representation ends in 1970, and this was done by design for heuristic reasons. This is because 18–20-year-olds were allowed to vote in the US after 1970, and these young voters tended to vote in lower proportions as compared with older Americans. Thus, the variable M_t changed in its nature after 1970. More specifically, M_t is a ratio, and after 1970 the denominator (total eligibles) got larger,

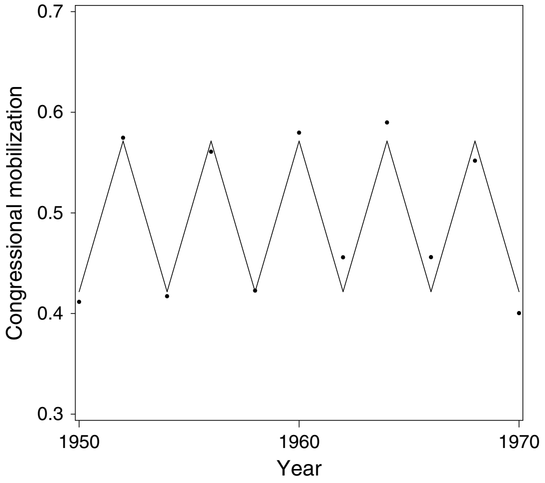


Figure 38.2 Congressional mobilization, 1950–70

but the numerator (total voters) did not grow equivalently. This means that the saw-toothed pattern found in Figure 38.2 continued after 1970, but at a lower overall level. In essence, this "dropped" or "bent" the pattern a bit after 1970, and a simple first-order linear difference equation model expressing finite oscillations could not capture this bend. A more sophisticated model—potentially a functionally nonlinear model—would have been able to capture this greater level of complexity, however.

A chaotic process differs from a periodic process in that changes in the values of the relevant variables never recur exactly, and this lack of repetition is not a consequence of a stochastic element. Chaotic processes often exhibit trajectories with dramatically diverse variations in variable values despite very small differences in initial conditions. Thus, chaotic processes lack periodic limit cycles, although variable variations typically "hover" within an identifiable neighborhood of an unstable equilibrium (a "strange attractor"), as I explain more thoroughly below.

Catastrophe processes experience sudden and dramatic changes that depart from previously existing dynamic processes. An earthquake is an obvious candidate for a phenomenon that can be modeled as a catastrophe process because incremental (i.e., regular process) changes in the positions of the tectonic plates eventually lead to a sudden departure from a previously established, pressure-defined equilibrium. In essence, a catastrophe process is such that the previously dominant equilibrium and its basin disappear, and a new equilibrium and its basin becomes dominant for the dynamical system. Examples of catastrophe models using social scientific data can be found in Brown (1995a, 1995b).

2 Competition and cooperation

Before going into greater depth with regard to chaos and catastrophe theories, it is worth describing the primary mechanisms by which nonlinear dynamics are commonly expressed in many models of social and political change. Algebraic formulations of social and political processes typically reference (either in isolation or combination) the ideas of competition and cooperation (see, e.g., Crosby, 1987). Most competition and cooperation formulations are expressed in terms of social systems.

Competitive processes are not goal seeking. Individuals and groups act selfishly to pursue their own interests in systems that are competitive in nature. Competitive systems are also nonlinear dynamically, in the sense that competing actors pursue locally-defined interests and goals. In a mathematical sense, this refers to the values of the partial derivatives of any particular system's Jacobian matrix. Such partials are always evaluated locally within a phase space, and their values typically change dramatically with respect to changes in the values of both the parameters and the state variables (as when one moves around in phase space). But competitive systems are also often (but not always) nonlinear in their functional form as well (in particular, see Brown, 2007).

Sometimes competitive systems can be hostile in nature, such as a nuclear arms race that threatens the survival of the planet. In other instances, such systems can be universally benign such that the overall environment within which the competition takes place is not harmed by the competition.

Cooperative systems are fundamentally different from competitive systems. Cooperative systems work to achieve collective goals. Individual actors within the system do not act without regard for the remainder of society. Thus, one can say that cooperative systems are goal seeking. In order for cooperative systems to achieve collective goals, it is necessary for elements of the system to be organized such that they are each dependent on one another. This results in a level of specialization among the various elements within society. Since there are collective goals, there must also be some centralization of power, which in turn implies some level of hierarchical control.

How does all of this relate to the subject of nonlinear dynamics? Both cooperative and competitive systems can experience nonlinear dynamics. However, cooperative systems have greater potential for experiencing linear dynamics. This is because it is possible for a cooperative system to have smooth (i.e., incremental) longitudinal change as one of its goals. Since a competitive system is locally dominated, nonlinear dynamics of any level of complexity can arise with great regularity. Within the social sciences, chaos and catastrophes are examples of highly nonlinear dynamics that can easily find residence within competitive systems.

Does this imply that cooperative social and political systems tend to express change as linear dynamics? The answer to this is unequivocally no, and this is due to the heterogeneous nature of most systems. In real life there are very few purely cooperative social systems. Most systems are hybrids with both competitive and cooperative elements. For example, a collective goal of reducing air pollution can be sought by society, but we may attempt to obtain this goal by instituting a competitive system involving “pollution permits” that are allocated to industries. Owners of some of the newer factories may be able to produce less pollution than their allotment, thereby allowing them to sell their excess permits to the owners of older factories who find it cheaper to buy such permits than to renovate their facilities. Such hybrid regulatory practices are called “marketplace” solutions because they allow factories to compete with one another while maintaining an eye on the overall level of allowed pollution.

3 Chaos theory

Chaos theory refers to a type of behavior with nonlinear dynamics that is both irregular and oscillatory. The study of chaos theory has assumed enormous importance in the physical and natural sciences, and it is now increasingly investigated with respect to the social sciences. For example, a recently published seminal volume edited by Kiel and Elliott (1996) presented a large variety of papers addressing both the theory and application of chaos theory in the social sciences. Chaos is encountered mathematically with certain sets of nonlinear deterministic dynamic models in which patterns of overtime behavior are not repeated no matter how long the model continues to operate. Some discrete-time models using nonlinear difference equations are known to exhibit chaotic dynamics under certain conditions using only one equation. However, in continuous-time models, the possibility of chaos normally requires a minimum of three independent variables, which usually requires an interdependent system of three differential equations. This requirement for continuous-time models can be changed in special cases if the system also has a forced oscillator or time lags.

Chaos can occur only in nonlinear situations. In multidimensional settings, this means that at least one term in one equation must be nonlinear while also involving several of the variables. Since most nonlinear models (and nearly all of the substantively interesting ones) have no analytical solutions, they must be investigated using numerically intensive methods that require computers (see Brown, 2007). Since the dimensionality and nonlinearity requirements of chaos do not guarantee its appearance, chaotic behavior is typically discovered in a model through computational experimentation that involves finding variable ranges and parameter values that cause a model to display chaotic properties.

Chaotic processes are quite common in real physical systems, and such processes may also be common to social systems, even though our present ability to identify and model such processes is still developing. Discovering real chaotic processes in physical and social systems is often quite difficult since stochastic noise is nearly always present as well in such systems, and it is not easy to separate truly random behavior from chaotic behavior. Nonetheless, mathematical tools continue to be developed that are aimed at sorting out these processes and issues.

Chaos has three fundamental characteristics. They are (a) irregular periodicity, (b) sensitivity to initial conditions, and (c) a lack of predictability. These characteristics interact within any one chaotic setting to produce highly complex nonlinear variable trajectories. Irregular periodicity refers to the absence of a repeated pattern in the oscillatory movements of the chaotically driven variables. Because of the irregular periodicity, Fourier analysis, graphing techniques, and other methods are commonly used to build a case for identifying chaotic processes (see Brown, 1995a).

A nonlinear model that has been among the most well studied with regard to chaos in discrete settings is a general form of a logistic map, and its chaotic properties were initially investigated by May (1976). This general logistic map is $Y_{t+1} = aY_t(1 - Y_t)$. Under the right conditions, this map can produce the standard S-shaped trajectory that is the trademark of the logistic process. However, oscillations in the trajectory occur when the value of the parameter a is sufficiently large. For example, when the value of parameter a is set to 2.8, the trajectory of the model oscillates around the equilibrium value of $Y_{t+1} = Y_t = Y^*$ while it converges asymptotically toward this equilibrium limit. But when the value of the parameter a is set equal to, say, 4.0, the resulting longitudinal trajectory never settles down toward the equilibrium limit, and instead continues to oscillate irregularly around the equilibrium in what seems to be a random manner that is caused by a deterministic process.

Figure 38.3 is a time series plot of the logistic map with parameter a set to 2.8. Note how the value of the variable Y_{t+1} settles down to an equilibrium point. This type of nonlinear behavior is oscillatory and convergent. Another way of looking at this behavior is with a stair-step diagram, and this is done in Figure 38.4. (See Brown, 1995a, for a discussion of how stair-step diagrams are constructed and interpreted.) Again note the convergence to the equilibrium point as the values of the first iterates become equal.

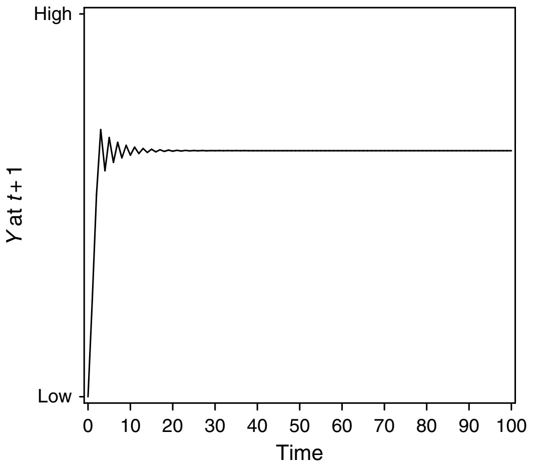


Figure 38.3 Time series of logistic map without chaos

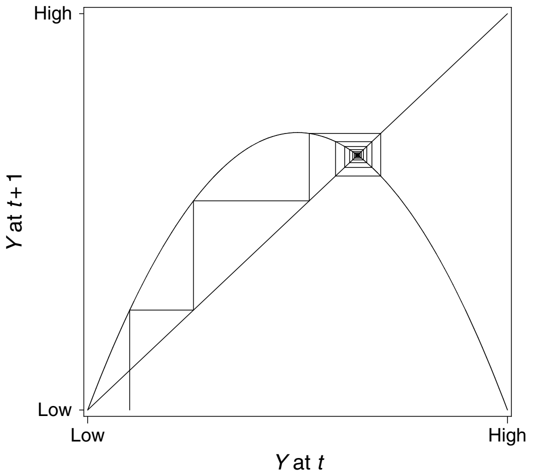


Figure 38.4 One-dimensional stair-step diagram of logistic map without chaos

Setting parameter $a = 4$, the time series plot changes dramatically with the emergence of chaos. This is seen in Figure 38.5. Time series plots are very hard to interpret in the presence of chaos. For this reason, stair-step diagrams are much more useful when chaos exists. Figure 38.6 is a stair-step diagram of the logistic map with parameter $a = 4$. From Figure 38.6 it is clear that the first iterates are not converging, which means that no single stable equilibrium exists for the system. Another way of saying this is that there is no convergence (periodic or otherwise) to a steady state. The discovery that such a simple model could exhibit chaotic properties was revolutionary to the subject on nonlinear dynamics. It implies that chaos may be quite common in nature, and human behavior would not be exempt from this.

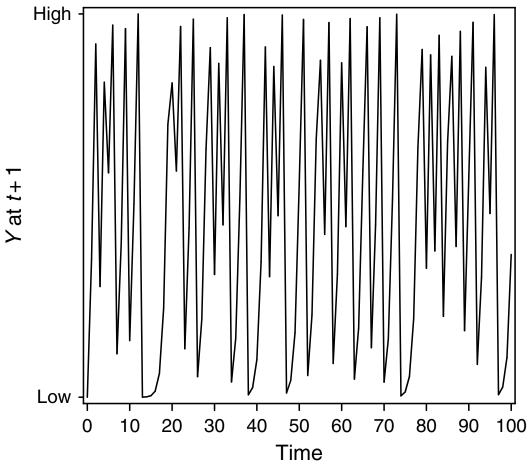


Figure 38.5 Time series of logistic map with chaos

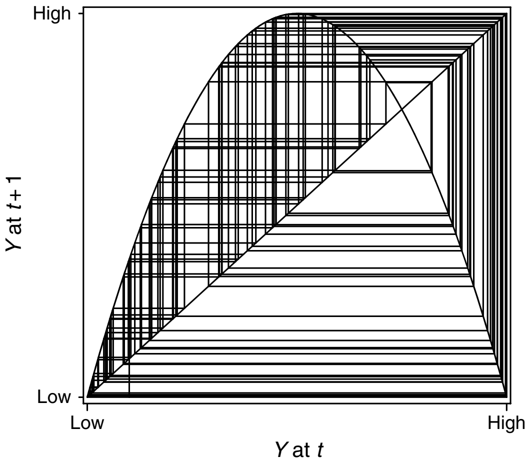


Figure 38.6 One-dimensional stair-step diagram of logistic map with chaos

The most famous continuous-time model that exhibits chaotic behavior is the so-called “Lorenz attractor” (Lorenz, 1963). This model is an interdependent nonlinear system involving three first-order differential equations, and it was originally used to analyze meteorological phenomena. The three equations are presented here as equations (2), (3), and (4).

$$dx/dt = s(y - x) \tag{2}$$

$$dy/dt = rx - y - xz \tag{3}$$

$$dz/dt = xy - bz \tag{4}$$

In this system, the three state variables are x, y, and z, whereas s, r, and b are parameters of the model. Most investigations of this system hold the parameters s and b constant while varying the value of r. The parameter r can be any positive number, but when $r < 1$, the origin is globally attracting, which means that there are no other competing attractors. When $r > 1$, there are three zero vectors that are found by setting the three derivatives in equations (2), (3), and (4) equal zero and then solving for the resultant state variables. However, now the origin is no longer a stable equilibrium point. Rather, all nearby trajectories in the system’s three-dimensional phase space now move away from the origin. When the parameter r is set at certain values, the other two zero vectors form what is called a “strange attractor.” The appearance of a strange attractor is a characteristic of chaos in such settings. In this case, the strange attractor forms a basin of attraction that draws trajectories into its neighborhood. But once trajectories arrive in the neighborhood of the strange attractor, the competing basins associated with each of the non-origin zero vectors interact to form an unstable area of phase space. In this area, trajectories forever move from one basin to the other and then back again, each time orbiting one zero vector before moving off to re-orbit the other zero vector. Moreover, the orbits never settle down into a periodic limit cycle. Using values suggested by Lorenz, we can set $s = 10$ and $b = 8/3$. Now, if r is set to 28, then the Lorenz system displays chaos. This situation can be seen in Figure 38.7.

There are many ways to analyze chaos. Some methods work with systems of equations that produce chaos. Other methods work with historical data by trying to discern if a deterministic chaotic process is present in combination with authentic random noise. Since chaos essentially mimics random noise, finding chaos in any body of data is a bit of a delicate art. Nonetheless, sophisticated methods do exist that accomplish just this. A review of some of these methods can be found in Brown (1995a). A useful discussion on how to use Lyapunov exponents while searching for chaos in small data sets can be found in Rosenstein, Collins and De Luca (1993); see also Hilborn (1994).

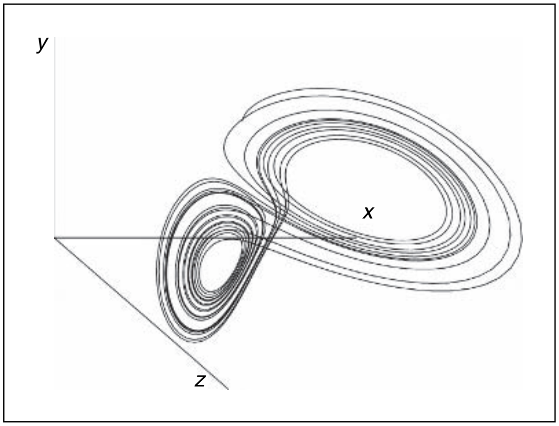


Figure 38.7 The strange attractor of the Lorenz system

Nonlinear models with forced oscillators are sometimes good candidates for exhibiting chaotic or near chaotic (i.e., seemingly random) longitudinal properties. In the social sciences, such a model has been developed and explored by Brown (1995b, Chapter 6). This model is a nonlinear system of four interdependent differential equations that utilizes a forced oscillator with respect to a parameter specifying alternating partisan control of the White House (or other relevant governmental institution). The dynamics of the system are investigated with regard to longitudinal damage to the environment, public concern for environmental damage, and the cost of cleaning up the environment. Variations in certain parameter values yield a variety of both stable and unstable nonlinear dynamic behaviors, including behaviors that have apparent random-like properties typically associated with chaos.

4 Catastrophe theory

Catastrophe theory refers to a type of behavior among some nonlinear dynamic mathematical models that experience nonlinear dynamics such that sudden or rapid, large-magnitude changes in the value of one variable is a consequence of a small change that occurs in the value of a parameter (called a "control parameter"). In this sense, catastrophe theory can model phenomena that loosely follow a "straw that broke the camel's back" scenario, although catastrophe theory can also be very general in its application. The modern understanding of catastrophe theory has its genesis in work by Thom (1975).

Nearly all early work with catastrophe theory employed polynomial functions in the specification of differential-equation mathematical models. In part, this was an important consequence of the generality of Thom's findings. Because all sufficiently smooth functions can be expanded using a Taylor series approximation (which leads us to a polynomial representation of the original model), it is possible to analyze the polynomial representation directly (see Saunders, 1980, p. 20). However, scientists can avoid using one of Thom's canonical polynomial forms by working with their own original theory-rich specifications as long as it is clear that the original specification has catastrophe potential (Brown, 1995a).

Bifurcations are fundamental to catastrophe theory. A bifurcation is an event that occurs in the evolution of a dynamic system in which the characteristic behavior of the system is transformed. With catastrophe theory, this occurs when an attractor in the system changes in response to change in the value of a parameter (called a "control parameter," because its value controls the manifestation of the catastrophe). A catastrophe is one possible consequence of a bifurcation (as compared with phenomena associated with, say, subtle bifurcations or explosive bifurcations).

The characteristic behavior of a dynamic system is determined by the behavior of trajectories, which are the values of the variables in a system as they change over time. When trajectories intersect with a bifurcation, they typically assume a radically different type of behavior as compared with that which occurred prior to the impact with the bifurcation. Thus, if a trajectory is "hugging" close to an attractor or equilibrium point in a system and then intersects with a bifurcation point, the trajectory may suddenly abandon the previous attractor and "fly" into the neighborhood of a different attractor. The fundamental characteristic of a catastrophe is the sudden disappearance of the influence of one attractor and its basin, combined with the dominant emergence of another attractor. Because multidimensional surfaces can also attract (together with attracting points on these surfaces), these gravity centers within dynamical systems are referenced more generally as attracting hypersurfaces, limit sets, or simply attractors.

A well-known example of the catastrophe specification that can easily be adapted to applications in the social sciences is the "spruce budworm problem." The model addresses the budworm that eats the foliage of balsam fir trees, and a full description of this model can be found in Ludwig, Jones and Holling (1978). The basic idea is that the budworm population grows at a logistic rate up to some equilibrium level that is determined by the limits of the food supply. But the specification can be adapted to include a predation term that allows for a sudden change in the budworm population. This predation often takes the form of an avian consumer of budworms.

To make this discussion more general, we can drop the budworm application entirely and consider the growth of any variable that has a logistic component combined with a loss term that contains the algebraic specification of predation as suggested above. Indeed, this has been done with respect to modeling marriages by Gottman, Murray, Swanson, Tyson and Swanson (2002, pp. 74–89). Such a model can be written as equations (5) and (6).

$$dN/dt = rN(k - N) - p(N) \tag{5}$$

where,

$$p(N) = BN^2/(A + N^2) \tag{6}$$

In equation (5), the first term on the right-hand side expresses logistic growth of the population N. The second term expresses loss to this population, which is more fully specified in equation (6). From equation (6), in the limit, $p(N)$ grows to B/A. But equation (6) acts as a switch since it allows loss (e.g., predation) to occur rapidly at values of N that approximately equal $\sqrt{A}$. Substituting equation (6) into equation (5), and then setting equation (5) equal to zero allows one to solve for the steady states of the model (i.e., the equilibria, N^*). But this formulation has a cubic term in N, which results in three equilibria for certain combinations of the parameter values. In such situations, one of the equilibria is unstable, while the other two are stable. The population dynamics are such that the value of N can be drawn toward one of the two stable equilibria while the value of one of the parameters (called a "control parameter" since its value can vary incrementally) changes. At some point, the ability of the first of the stable equilibria to continue to "hold on" (i.e., continue to influence) the value of N vanishes, and N is quickly captured within the basin of the other attracting equilibrium. This can produce a "cusp" catastrophe.

Another model, this time due to Zeeman (1972), illustrates a simple cusp catastrophe with yet a different specification. This model is used to describe the change in muscle fiber length (variable x) in a beating heart. The control parameter A (which in this instance refers to the electrochemical activity that ultimately instructs the heart when to beat) can change in its value continuously, and it is used to move trajectories across an equilibrium hypersurface that has catastrophe potential. The parameter q identifies the overall tension in the system, and f is a scaling parameter. The two differential equations in this system are $dx/dt = -f(x^3 - qx + A)$, and $dA/dt = x - x_1$. Here, x_1 represents the muscle fiber length at systole (the contracted heart equilibrium). Setting the derivative $dx/dt = 0$, we will find between one and three values for x, depending on the other values of the system. When there are three equilibria for x for a given value of the control parameter A, one of the equilibria is unstable and does not attract any trajectory. The other two equilibria compete for the attention of the surrounding trajectories, and when a trajectory passes a bifurcation point in the system, the trajectory abandons one of these equilibria and quickly repositions itself into the neighborhood (i.e., the basin) of the other equilibrium. This rapid repositioning of the trajectory is the catastrophe.

A sample picture of a catastrophe surface is shown in Figure 38.8. In this figure, note how the trajectories move to one of the two stable equilibria in the upper left or lower right of the graph. Note also that when a trajectory moves over the lip of the cusp equilibrium surface, it experiences a rapid vertical shift as it is drawn toward the other stable equilibrium. This rapid vertical shift is the classic signature of the catastrophe phenomenon.

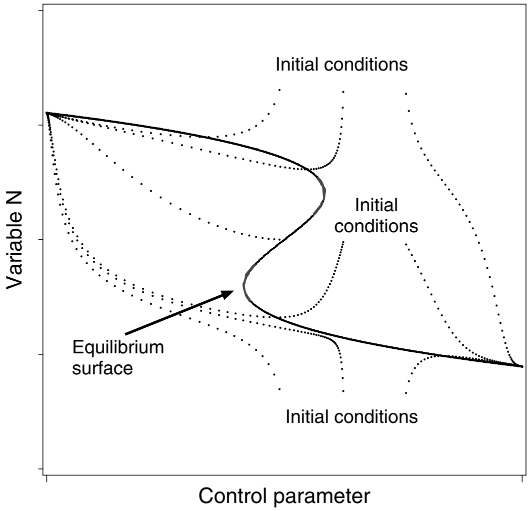


Figure 38.8 A sample "cusp" catastrophe equilibrium surface with associated variable dynamics

Two social scientific examples of nonlinear differential equation dynamic catastrophe specifications have been developed and explored by Brown (1995b, Chapters 3 and 5). One example involves the interaction between candidate preferences, feelings for a political party, and the quality of an individual's political context or milieu during the 1980 presidential election in the United States. The other example addresses the interaction between the partisan fragmentation of the Weimar Republic's electorate, electoral deinstitutionalization, and support for the Nazi party. Both examples are fully estimated; the first uses both individual and aggregate data while the second employs aggregate data only.

5 The future of nonlinear modeling in the social sciences

It is highly probable that the study of nonlinear dynamics—and in particular the use of functionally nonlinear continuous and discrete-time mathematical models—will continue to grow in the social sciences. Minimally, the limitations of the linear regression model will almost certainly force social scientists to continue to expand their use of mathematical techniques into areas that have traditionally been exploited more heavily in the natural and physical sciences. It is not that social scientists should stop using linear models. It is just that growth and exploration are natural to the scientific enterprise, and research into nonlinearity promises to be an area of high yield in terms of scientific productivity in the social sciences. Also, new languages of mathematical modeling—such as graph algebra (Brown, 2007b)—are likely to ease the mechanics of nonlinear model building just as the use of such languages have similarly assisted the field of engineering. It is difficult to predict with certainty how any field will develop. But it seems highly likely that the drive to understand the complexity of human societies will lead to the exploitation of mathematical models that similarly address greater levels complexity. The study of nonlinearity is one of a number of competing pathways to this richer level of understanding.

Glossary

Bifurcation A point at which there is a qualitative change in the longitudinal behavior of a dynamic system or variable.

Catastrophe theory A theory originally developed by the French mathematician, René Thom, that describes sudden and large magnitude change in one variable as a consequence of incremental change in the value of a control parameter.

Chaos A branch of mathematics in which change in the value of one or more variables in a deterministic model appears random in nature, yielding great sensitivity to initial conditions.

Nonlinear dynamics Longitudinal change that is not defined in terms of a constant increment over time, usually associated with highly nonlinear dynamic model specifications.

Phase space The dimensions housing change in a system's dependent variables such that time (an independent variable) is suppressed.

Trajectory Sequential change in one or more dependent variables within a system, usually within the context of phase space.

References

Brown, C. (1991). *Ballots of Tumult: A Portrait of Volatility in American Voting.* Ann Arbor: University of Michigan Press.

Brown, C. (1995a). *Chaos and Catastrophe Theories.* Thousand Oaks, CA: Sage.

Brown, C. (1995b). *Serpents in the Sand: Essays on the Nonlinear Nature of Politics and Human Destiny.* Ann Arbor: University of Michigan Press.

Brown, C. (2007a). *Differential Equations: A Modeling Approach.* Thousand Oaks, CA: Sage.

Brown, C. (2007b). *Graph Algebra: Mathematical Modeling with a Systems Approach.* Thousand Oaks, CA: Sage.

Crosby, Robert W. (1987). Toward a classification of complex systems. *European Journal of Operational Research*, 30: 291–293.

Gottman, J. M., Murray, J. D., Swanson, C., Tyson, R. and Swanson, K. R. (2002). *The Mathematics of Marriage: Dynamic Nonlinear Models*. Cambridge: MIT Press.

Hilborn, R. C. (1994). *Chaos and Nonlinear Dynamics.* New York: Oxford University Press.

Hirsch, M. W. and Smale, S. (1974). *Differential Equations, Dynamical Systems, and Linear Algebra.* New York: Academic Press.

Hirsch, M. W., Smale, S. and Devaney, R. (2003). *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, 2nd edn. New York: Academic Press.

Kiel, L. D. and Elliott, E. (1996). *Chaos Theory in the Social Sciences: Foundations and Applications*. Ann Arbor, Michigan: University of Michigan Press.

Lorenz, E. N. (1963). Deterministic non-periodic flow. *Journal of Atmospheric Science*, 20: 130–141.

Ludwig, D., Jones, D. S. and Holling, C. S. (1978). Qualitative analysis of insect outbreak systems: The spruce budworm and the forest. *Journal of Animal Ecology*, 47: 315–332.

May, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature*, 26: 459–467.

Richards, D. (ed.) (2000). *Political Complexity: Nonlinear Models of Politics*. Ann Arbor, Michigan: University of Michigan Press.

Richardson, L. Fry (1960). *Arms and Insecurity*. Chicago: Quadrangle Books.

Rosenstein, M. T., Collins, J. J. and De Luca, Carlo J. (1993). A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D*, 65: 117–134.

Saunders, P.T. (1980). *An Introduction to Catastrophe Theory*. New York: Cambridge University Press.

Thom, René (1975). *Structural Stability and Morphogenesis*. Reading, MA: W.A. Benjamin.

Zeeman, E.C. (1972). Differential equations for the heartbeat and nerve impulse. In C.H Waddington (ed.), *Towards a Theoretical Biology*, Vol. 4, pp. 8 – 67. Chicago: Edinburgh University Press.

Index

Age, period, and cohort effects 87, 219–32
  dummy variable regression approach 223–26
Aggregate level analysis 4–6
Akaike information criterion (AIC) 470, 556–7, 563, 586–7
Allgemeinen Bevolkerungsumfragen der Sozialwissenschaften (ALLBUS) 35
Alternating least squares (ALS) 9, 334
ARIMA (*see* Time series: ARIMA)
Attrition, panel (*see* Missing data)
Autocorrelated error 206, 233, 237–8, 240, 242, 486–7, 494, 499, 510
Autocorrelated error models (*see* Time series: autoregressive error models)
Autocorrelation 250–4, 264, 320, 328, 373, 375, 380, 471, 583–5, 587–9, 590–2, 594, 602, 624–5
Autoregression, autoregressive models (*see* Time series)

Bayesian information criterion (BIC) 556–7, 563, 586–7
British Birth Cohort Studies (BCS) 68–74, 169, 176
British Household Panel Surveys (BHPS) 131, 170–2, 175–6
British Millennium Cohort Study 67–84
British Social Attitudes Panel 131

Catastrophe theory (models) 11, 660–2
Causal analysis/causal inference 4, 259–78, 279–93, 475–77
Census data 4, 13–31
Change in measurement 8, 42
Change in relationships among variables 4
Change, measuring 3, 42–44, 148–150, 279–83, 301–9
  qualitative 160–63, 281–2, 506–8
Chaos theory (models) 11, 657–60
Cholesky decomposition 206, 208
Cohort 93–94
  as a unit of analysis 221–23, 429–37
  effects (*see* Age, period, and cohort effects)
  studies 67–84, 85–106
Conditional change model 508–11, 517–19
Conditional logistic regression (*see* Logistic regression, conditional)
Confirmatory factor analysis 142–6
Configural frequency analysis 9, 313–32
Consumer Expenditure Survey (CES) 127
Cox regression (*see* Event history analysis: semiparametric)
Cross-lagged panel correlation 155, 263–4, 271
Cross-sectional analysis 346, 352, 354
Cross sectional data and designs 15, 31–2, 44–5, 49, 51, 80, 90–92, 96, 112–5, 124–5, 127–8, 130–1, 141, 150, 168, 173, 177, 179–80, 249–257, 398, 476–9, 493–5, 505–6, 509–10
Current Population Survey (CPS) 128–9

Developmental change, analysis and models 6, 8, 85–106, 153–65, 219–32, 299–300, 307–10, 318, 323–6, 342–53, 389–403, 444–458, 510, 523–544, 545–564, 565–8, 569–574, 639–40
Difference equations 654
Differential equation models 11, 639–52, 653–4, 657–62
Discrete-time survival analysis (*see* Event history analysis: discrete)
Dynamic analysis/models 8, 11, 249–57, 486–7, 499–501

Ergodic and nonergodic systems 249, 251–3, 256–7, 644
European Community Household Panel Survey (ECHP) 127, 169

Event history data and analysis 9–11, 63–4, 92–3, 288–92
discrete 10, 441–63
multistate 393–397
nonparametric 9, 389–403, 423–7
parametric 9, 421–39
recurrent events 397–402
semiparametric (Cox proportional hazards) 9, 405–419
Experimental research 9, 279–93
Experiments with survey forms 39–40

Factorial invariance 142–6, 156, 162–3
Fixed effects models 188, 207, 213, 234–44, 256, 272–6, 469, 473, 479–81, 505, 513–5, 517, 549, 551, 557–9, 561, 569

General Health Questionnaire (GHQ) 130–1
General Social Survey (GSS) 6, 33–48, 362, 370
Generalized estimating equations (GEE) 10, 285–6, 290–2, 467–74
German Life History Study (GLHS) 85–106
German Socioeconomic Panel (SOEP) 127, 130, 171–2, 476
Granger causality 245, 595–7, 634
Graphical techniques 8, 199–217
Growth curves (*see also* Latent growth curve analysis; Multilevel growth curve model) 191, 332–6

Hausman test 235–6, 273, 483
Health and Retirement Survey (HRS) 125, 172, 176
Health Surveys for England (HSE) 131
Heterogeneity, analysis of 236–7
Hierarchical linear modeling (HLM) (*see* Multilevel modeling)
Historical change, analysis and models 8, 27–8, 219–32, 409–15, 586–92, 608–9, 613–15, 617–18, 629–36
Household Income and Labour Dynamics in Australia survey (HILDA) 170–2

International Social Survey Program (ISSP) 33, 35–9, 44–4
Interrupted time series analysis (*see* Time series: interrupted)
Intervention analysis (*see* Time series: interrupted)
Intraindividual change (*see* Developmental change, analysis and models)

Lagged effect
Lagged endogenous variable models 237–8, 240–5, 264, 267–73, 286–8, 343–4, 348, 353–5, 486–501, 509–10, 519, 580, 594–5, 599, 623–4, 626–36
Lagged predictors/lagged effects 238, 263, 265–270, 273, 417, 517, 519, 571–3, 594–99, 623–4
Lagged residual 237
Latent differential equations 644–41
Latent change analysis 354
Latent class analysis 9, 357–71, 373–85
Latent growth curve analysis 10–11, 129, 132, 135–6, 143–5, 153–61, 191–5, 264, 273–6, 373, 377–9, 478, 523–44, 565, 579, 599, 644–6
Latent variable 10–11, 129, 132, 135–6, 143–5, 155–8, 162, 354, 373–83, 477, 495–502, 526–8, 545, 643–50
Life course perspective and research (*see* Developmental change, analysis, and models)
Life history calendar 92, 116, 118–20
Linear panel analysis (*see* Panel analysis, linear)
Logistic regression 59–64, 281–2, 288–91, 375, 380–82
conditional 512–4
in event history analysis 453–61
multilevel 565–76
panel analysis 10, 505–522
Loglinear models 317–9
Longitudinal panel design 6, 67–84
Locally weighted regression (lowess) 203, 204, 206, 209–10, 402, 409

Mahalanobis distance 207, 209
Markov models 373–385, 394–397
Millennium Cohort Study 67–83
Missing data (*see also* Panel attrition/retention; Panel missing data) 8, 55–61, 95–6, 185–96, 376
initial nonresponse 34, 59
mean imputation 185
missing at random (MAR) 185–8
missing completely at random (MCAR) 186
missing not at random (MNAR) 186, 188–90
multiple imputation 186
pattern mixture models 189–90
selection models 188–9
MIT Growth and Development Study 200
Mixture latent markove model 373–9
Monitoring the Future Survey (MTF) 128

Moving average (MA) model (*see* Time series: moving average (MA) model)
MRC National Survey of Health and Development, 1946 cohort 68–72
Multiple correspondence analysis 335–42
Multilevel modeling 10–12, 276, 284–6, 289–292, 333, 381, 490, 502, 643
  and Cox model 417
  and latent class model 375, 378
  multilevel change model 10, 565–76
  multilevel growth curve model 10–11, 517, 520–1, 525–6, 545–64, 579, 599
  multilevel logistic regression (*see* Logistic regression: multilevel)
Multivariate analysis of variance (MANOVA) 284–6, 290

National Child Development Study (NCDS) 68–74
National Crime Survey (NCS) (*see* National Crime Victimization Survey)
National Crime Victimization Survey (NCVS) 6, 10, 49–65, 128
National Data Program for the Social Sciences (NDPSS) 33
National Election (Panel) Study (NES) 476, 488, 491
National Health Interview Survey (NHIS) 124–5
National Longitudinal Study of Youth (NLSY) 117, 172, 546–8
National Opinion Research Center (NORC) 33–34, 42
National Youth Survey (NYS) 114–5, 128, 379–81, 569, 505–606, 513, 516–7, 569
Netherlands Schoolproject 348–53
Nickell bias 238
Nonexperimental research 9
Nonlinear dynamics 10, 653–7, 663–4
Nonrecursive models 269–71
Nonresponse, initial (*see* Missing data: initial nonresponse)
Nonstationarity (*see* Time series: stationarity)
Number of cases and periods 11–12

Observational data 297–311
Optimal scaling techniques 9, 333–56
Ordinary least squares (OLS) 233–4, 621–2
Oregon Youth Study 155–6, 158–160, 528–9

Panel analysis, linear 10–11, 260–78, 475–504
Panel analysis, logistic regression 505–522
Panel attrition/retention 8, 49, 51, 59, 63, 72–74, 80–81, 167–84, 185
  frequency of interviewing 170–1
  incentives 179–81
  interview length and complexity 171
  Keep In Touch Exercises (KITEs) 174–5
  mixed mode data collection 171–2
  tracing/tracking 173–7
Panel conditioning 8, 123–38
Panel corrected standard errors (PCSEs) 238–4, 273
Panel heteroskedasticity 238–40
Panel missing data 185–196
Panel Study of Income Dynamics (PSID) 169–72, 175–6, 476
Period effects (*see* Age, period, and cohort effects)
Pooling cross-sectional and time series data 8, 233–48, 271–3, 475, 483, 516
  optimal scaling approaches 342–53
Population averaged model 290, 469–70, 571–2
Proportional hazards model (*see* Event history analysis: semiparametric)

Qualitative analysis 9, 12, 297–311
Qualitative data 9, 12, 260–3, 288, 297–311

Random coefficient model (*see* Multilevel modeling)
Random effects/random effects models (*see also* Unit-specific model) 188, 191, 205, 213, 234–6, 246, 272–4, 378, 416–7, 469, 473, 482–3, 505, 549–555, 565, 568, 573
Recall, length of 8, 96–101, 109–121
Refusal conversion 179
Regression fallacy 149–50
Regression to the mean 146–7, 280
Reliability 7–8, 34–35, 40, 139–51, 153–65
Repeated cross-sectional design/data 5, 33–48, 276–7, 373–379, 621–32
Replication 34–35, 42–43
Residuals, analysis of 8, 204–215, 407–413
  cumulative sums of residuals 210–215
  moving sums of residuals 211–215
  supremum test 211–215
Retrospective designs 6–8, 85–106, 109–121
Revolving (or rotating) panel design 6, 49–65, 124, 169

Samples/Sampling 4, 7, 26, 49–50, 75–6, 91–2, 169–73
Schwartz-Bayes criterion (SBC) 586
Shared method variance 156

Spectral analysis (*see* Time series: Spectral analysis)
Stability in measurement (*see* Reliability)
Structural equation modeling 10–12, 155–63, 249–56, 264–76, 442, 477–8, 487–501, 520–1, 523, 544, 596, 644–50
Subject-specific model (*see* Unit-specific model; Random effects)
Survey of Income and Program Participation (SIPP) 127, 169–71
Survey research 33–48, 49–65, 67–84, 168–73
Survival analysis (*see* Event history analysis)

Telescoping 52, 111, 129
Temporal order 3, 49
Time-ordered cross-sectional design 3
Time plots 200–204
Time series: 4, 11, 16, 23–4, 153, 233–48, 250, 323, 354, 475, 486, 643–8, 655–60
  analysis 11, 579–600, 601–619, 621–637
  ARIMA (Autoregressive integrated moving average) 579–600, 624–8
  autoregressive error models 238, 250–6, 286–8, 594–6, 599, 622–4, 630–33
  autoregressive (AR) model 242, 579, 584–94, 641–3
  comparison of models 621–637
  cross-spectrum and bivariate spectral analysis models 615–8
  differencing (*see* Time series: integrated (I) models)
  Dickey–Fuller tests 239–40, 585
  Durbin–Watson test 630–31
  forecasting 597–8
  integrated (I) models 584–5, 625, 636–7
  intervention (*see* Time series: interrupted)
  interrupted 593–4, 626–36
  lagged endogenous variable models (*see* Lagged endogenous variable models)
  moving average (MA) model 579–94, 613
  multivariate 594–9, 622–3
  optimal scaling models 353–4
  prewhitening 596–7, 624–6
  random shock 580, 583–4
  spectral analysis 11, 601–619
  stationarity 239, 579–80
  time series regression 343, 347–8, 621–3
  transfer function 595–9
  vector autoregression 595–7, 599
Time-varying predictors/covariates 10, 373–9, 405, 423, 425, 444, 447, 459–60, 519, 548, 560–1, 568, 570, 575, 579–80, 599
Total population design 4
Transitions, analysis of 328–9, 373–84, 389–403
Tracing (tracking) respondents 174–7
Trends 23–4, 43–4, 239, 242–4, 322–3 (*see also* Time series)

Unconditional change model 508–511, 514–7
Uniform Crime Reports (UCR) 4
Unit-specific model (*see also* Random effects) 246, 290, 469–70, 483–7, 499, 511–3, 565, 569, 572–4
Unobserved heterogeneity 478–479, 499–501

Working correlation matrix 470–4
Wu-Hausman test (*see* Hausman test)

Youth Attitude Tracking Survey (YATS) 131–2
Youth In Transition study 266

Zentrum für Umfragen, methoden, und Analysen (ZUMA) 35, 87